STATISTICAL ISSUES IN COMBINING MULTIPLE GENOMIC STUDIES: QUALITY ASSESSMENT, DIMENSION REDUCTION AND INTEGRATION OF TRANSCRIPTOMIC AND PHENOMIC DATA

by

Dongwan Don Kang

BA, Seoul National University, Korea, 2000

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Dongwan Don Kang

It was defended on

May 20th, 2011

and approved by

Dissertation Advisor:
George Tseng, ScD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Naftali Kaminski, MD
Professor
Department of Medicine
School of Medicine
University of Pittsburgh

Committee Member:
Lisa Weissfeld, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Michael Barmada, PhD
Associate Professor
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

STATISTICAL ISSUES IN COMBINING MULTIPLE GENOMIC STUDIES: QUALITY ASSESSMENT, DIMENSION REDUCTION AND INTEGRATION OF TRANSCRIPTOMIC AND PHENOMIC DATA

Dongwan Don Kang, PhD

University of Pittsburgh, 2011

Genomic meta-analysis has been applied to many biological problems to gain more power from increased sample sizes and to validate the result from an individual study. As for the study selection criteria, however, most literatures depend on qualitative or ad-hoc numerical methods, and there has not been an effort to develop a rigorous quantitative evaluation framework. In this thesis, we proposed several quantitative measures to assess the quality of a study for a meta-analysis. We have applied the proposed integrative criteria to multiple microarray studies to screen out inappropriate studies and also confirmed the necessity of proper exclusion criteria using real meta-analyses. By simulation studies, we showed the effectiveness and robustness of the proposed criteria. Secondly, we have investigated simultaneous dimension reduction frameworks for down-stream genomic meta-analysis. Currently, most microarray meta-analyses focus on detecting biomarkers; however, it is also valuable to seek a possibility of meta-analysis in unsupervised or supervised machine learning, particularly dimension reduction when multiple studies are combined. We proposed several simultaneous dimension reduction methods using principal component analysis (PCA). Using five examples of real microarray data, we showed the information gain obtained by adopting our proposed procedures in terms of better visualization and prediction accuracy. In the third component, we pursued a novel approach to elucidate undefined disease phenotypes between interstitial lung disease (ILD) or chronic obstructive pulmonary disease (COPD). By applying unsupervised learning technique to both

clinical phenotypes and gene expression data obtained from well characterized large number of cohort, we successfully showed the existence of intermediate phenotypic group who have both disease characteristics and divergent phenotypes in clinical and molecular features. Public health importance of our findings is that we showed current clinical definitions and classification do not account for the large number of patients having intermediate phenotypes or less common features that are often excluded from clinical trials and epidemiology reports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Microarray is now de facto standard technology to reveal gene expression of tens of thousands genes simultaneously. Since its introduction, it has been used to generate tremendous amount of data that are accumulated in the public repositories such as NCBI Gene Expression Omnibus [18], EBI ArrayExpress [76] and Stanford Microarray Database [95]. Despite the huge amount of available data, the analysis of individual microarray study often suffers from limited statistical power caused by small sample sizes and inconsistent result with other related studies due to heterogeneous cohorts, data annotation or preprocessing errors [21, 105, 44, 16].

Meta-analysis has recently gained popularity in the genomic research in light of its successful application to traditional epidemiological and medical researches. From a recent review paper, a total of 383 papers related to microarray meta-analysis have been published since 2004 to end of 2010. Many microarray meta-analysis methods have been proposed to increase statistical power for differentially expression gene detection using Fisher's method [25, 86], LASSO method [30], random effects model [9, 101], Bayesian methods [49, 108], rank-based methods [6, 38], and other methods [74, 64]. Recently, a statistical framework for microarray pathway meta-analysis was also proposed [94]. Hong [37] and Campain [7] compared performance of different microarray meta-analysis methods and Ramasamy [84] discussed key issues and a practical guide for performing microarray meta-analysis.

Although the advantages of genomic meta-analysis are apparent, a quantitative and systematic approach to decide the inclusion or exclusion criteria of microarray studies for a meta-analysis has not been pursued yet. Instead, most literatures depend on subjective expert opinions or ad hoc criteria (e.g. microarray platforms, tissue types used or number of sample sizes) [87, 35, 113, 68, 97, 17]. Conceptually, including a bad quality or outlying

study in the information integration can greatly dilute strength of signal, decrease statistical power or even distort final biological conclusions. To alleviate such potential pitfalls in meta-analysis [23], it is necessary to develop objective inclusion/exclusion criteria.

In chapter 2, we propose quantitative measures to assess the quality and consistency of microarray studies for meta-analysis. Specifically, we developed six quality control measures and utilized principal component analysis (PCA) biplots and an averaged rank summary score to assist selection of studies. We then applied the proposed methods to four examples, including brain cancer, prostate cancer, idiopathic pulmonary fibrosis (IPF) and major depressive disorder (MDD). Impacts and effectiveness of the proposed inclusioin/exclusion criteria on the final meta-analysis results in real examples are evaluated. Additional simulations were also performed to show the robustness and effectiveness of proposed methods.

Another issue that current genomic meta-analysis literatures are lack of is the possibility of down-stream statistical analysis other than usual biomarker detection or gene set enrichment analyses. When it comes to the other down-stream analysis, we will investigate in chapter 3 the simultaneous dimension reduction of multiple microarray studies using Principal Component Analysis (PCA) by finding a common PC subspace, named as MetaPCA.

The PCA is one of the most popular techniques that enable us to explore high dimensional data space through low dimensional projection. Each Principal Component (PC) is orderly derived to be a linear combination of original features in that it explains as much of the variance as possible. The first PC is chosen to have the largest variance, and the next PCs are sequentially derived to have the largest variance in the orthogonal subspace of selected PCs [46].

One of the many applications of PCA is to visualize high dimensional subjects in two or three dimensional PC subspace. The derived subspace is the optimal choice to represent as much information (variance) as possible in such a reduced dimension, i.e. it minimizes the sum of squares of the projection errors. In many cases, this approach is very successful [85, 92, 63, 11, 120]. In the other hand, Sparse PCA was recently proposed to attain a better interpretation of each PC—it gives sparse loadings so that each PC can be interpretable by a smaller set of original features. It also can be served as a good technique for an unsupervised feature selection [47, 67, 123, 13, 115, 48, 60].

Another possible application of PCA is to utilize PCs as an input for other statistical methodologies such as regression analysis or various multivariate techniques [46, 39, 53]. The two main advantages of PC based approaches are to overcome the multicollinearity problem which occurs when highly correlated variables are included together in an analysis and to alleviate the curse of dimensionality which numerous multivariate methods are suffered from [4, 2].

When PCA is applied to microarray data, it is hoped that the chosen PCs elucidate the informative low dimensional manifold such as interesting patterns or clusters of subject or genes. However, microarray data is noisy, and PCA is prone to outliers; often time we fail to get an effective representation in the reduced dimension. Moreover, when we have several similar studies and try to compare several PCA results in parallel, each PC subspace is not comparable. An intuitive approach is to project new data sets to the given PC subspace. However, this approach is not efficient in that it fails to utilize all information from available data sets and not robust in that it depends totally on the quality of target PC subspace. Better approaches should be the ones that can use all information and is robust such that the driven PCs retain only informative shared subspace and reveal the true separation among subjects.

In chapter 3, we propose two approaches that resolve the issues occurred by an individual PCA. We hypothesize that several PCA results from compatible data sets have an advantage to elucidate the common cause of variance throughout data sets regardless of individual noise and heterogeneity of each study. Focusing on the commonness among studies is the philosophy of what we are trying to do with MetaPCA by finding the "optimal common subspace" in multiple studies.

Although some similar ideas have been in the literature, in our knowledge there was no effort to apply the concept of MetaPCA to genomic researches. Current literatures regarding to genomic meta analysis are wholly focused on biomarker detection or gene set enrichment analyses. Here, we have investigated the simultaneous dimension reduction of multiple microarray studies using MetaPCA, and we show the usefulness of common subspace in terms of dimension reduction for data visualization and classification.

As the last aim, in chapter 4 we have investigated methods to integrate transcriptomic and phenomic data. It is necessary to define phenotypes of disease states to investigate their underlying molecular mechanisms and develop new treatment strategies, therapeutics, and biomarkers. Using chronic lung diseases which are commonly thought of as clinically distinct, we demonstrate a computational approach to"reverse phenotype" patients using both clinical and gene expression data. We acquired lung tissue, computed tomography, and clinical data on 474 subjects who were initially given a clinical diagnosis of either interstitial lung disease (ILD) or chronic obstructive pulmonary disease (COPD). We performed unsupervised clustering on patients with both phenomic and genomic data. We developed insightful feature visualization tools to explore and interpret the clusters. Pathway analysis and clinical feature correlation are performed to characterize and annotate the identified intermediate phenotypic patients. We showed the convergence/divergence patterns of disease phenotypes in the integrated clustering by clinical and molecular features. Large number of patients was in off-diagonal clusters which represent discordancy between clinical and molecular phenotypes. This is the first paper that systematically integrate genomic and phenomic data in ILD and COPD. We identified new clusters of intermediate phenotypic patients that may lead to improved understanding of these diseases and novel therapeutic approaches. Approximately 24 million adults in the US are affected by chronic lung diseases and 119,000 dies each year mostly due to COPD or ILD. Our findings reflect that current clinical definitions and classification do not account for the large number of patients having intermediate phenotypes or less common features that are often excluded from clinical trials and epidemiology reports.

Overall the theme of this thesis is information integration to maximize information gain and resolve pitfalls in individual or single type of data analysis. We developed sophisticated and objective methodologies in the numerical evaluation of study quality by adopting statistical inferences and computation approaches. We showed both the power of genomic meta-analysis and the necessity of proper inclusion/exclusion criteria. Based on the quality control results, in the second part, we selected a set of quality studies for a meta-analysis, and we could successfully observe the information gain through our proposed novel genomic integrative analysis in dimension reduction. One of conclusions in the analysis is that we

need homogeneous studies to borrow beneficial information from other studies. In the third part, our innovative integration approach of transcriptomic and phenomic data elucidated homogeneous clusters of patients who are not defined previously. Our work can be served as a framework for further studies that seek integrative interpretation of interesting patients group which may lead to consensus definition of novel subtype of the two chronic lung diseases.

In chapter 2, integrative quality control criteria are proposed, and application to real data sets and simulation studies are followed. Chapter 3 focuses on development of MetaPCA and its application to real data sets. In chapter 4, we present a novel framework to integrate phenomic or transcriptomic data. Finally, in chapter 5, we discuss overall conclusion of our work and future direction to extend our methodologies.

2.0 METAQC: OBJECTIVE QUALITY CONTROL AND INCLUSION/EXCLUSION CRITERIA OF GENOMIC META-ANALYSIS

Genomic meta-analysis that combines multiple microarray studies have been widely applied to increase statistical power and to validate results from individual studies. Currently, the inclusion/exclusion criteria to the analysis mostly depend on ad-hoc expert opinion or naïve decision by sample size or array platform. To our knowledge, no objective quality assessment has been developed. In this paper, we propose six quantitative quality control measures, covering internal homogeneity of co-expression structure among studies (internal quality control, IQC), external consistency of co-expression pattern with pathway database (external quality control, EQC), and accuracy and consistency of differentially expressed gene detection or enriched pathway identification (accuracy quality control, AQCg and AQCp; consistency quality control, CQCg and CQCp). Each quality control index is defined as the minus log transformed p-values from formal hypothesis testing. Principal component analysis biplots and a standardized mean rank are applied to assist visualization and decision. We applied the proposed method to four microarray meta-analysis examples: 7 brain cancer studies, 9 prostate cancer studies, 8 idiopathic pulmonary fibrosis studies, and 17 major depressive disorder studies. The identified problematic studies are scrutinized to identify technical and biological causes (e.g. sample size, platform or tissue processing) of their bad quality or irreproducibility to determine exclusion from the final meta-analysis. The results generated systematic suggestions to exclude problematic studies for genomic meta-analysis of microarray. The method can be extended to meta-analysis applications of genome-wide association studies or other sequence-based new technologies.

2.1 INTRODUCTION

Microarray has been an effective and economic technology to monitor gene expression of tens of thousands genes simultaneously. Since its introduction, it has been used to generate tremendous amount of data that are accumulated in the public repositories such as NCBI Gene Expression Omnibus [18], EBI ArrayExpress [76] and Stanford Microarray Database [95]. Despite the huge amount of data available, the analysis of individual microarray study often suffers from limited statistical power caused by small sample sizes and the results are inconsistent result with other related studies due to heterogeneous cohorts, data annotation or preprocessing errors [21, 105, 44, 16].

Meta-analysis has gained popularity in the genomic research in light of its successful application to traditional epidemiological and medical researches. Many microarray meta-analysis methods have been proposed to increase statistical power and obtain validated results across studies. Methods for detecting differential expression genes include Fisher's method [25, 86], Stouffer's method [102], LASSO [30], random effects model [9, 101], Bayesian methods [49, 108], rank-based methods [6, 38], and others [74, 64]. In addition to DE gene detection, a statistical framework for microarray pathway meta-analysis was also proposed [94]. Hong et al. [37] and Campain and Yang [7] compared performance of different microarray meta-analysis methods. Ramasamy et al. [84] discussed key issues and a practical guide for performing microarray meta-analysis. Although the advantages of genomic meta-analysis are apparent, a quantitative and objective approach to decide the inclusion or exclusion criteria of microarray meta-analysis has not been pursued yet, to our knowledge. Instead, most literatures depend on subjective expert opinions or ad hoc criteria (e.g. platforms, tissue types used or number of sample sizes) [87, 35, 113, 68, 97, 17]. Conceptually, including a bad quality or outlying study in the information integration can greatly dilute information contained, weaken statistical power or even distort final biological conclusions. To alleviate such potential pitfalls in meta-analysis [23], it is necessary to develop an objective inclusion/exclusion evaluation tool.

In this paper, we proposed quantitative measures to assess the quality and consistency of microarray studies for meta-analysis. Specifically, we developed six quality control mea-

sures (see Table 2 for a brief summary) and utilized principal component analysis (PCA) biplots and a standardized mean rank (SMR) summary score to assist identification of problemtic studies. We then applied the proposed methods to four examples, each containing 7 brain cancer, 9 prostate cancer, 8 idiopathic pulmonary fibrosis (IPF) and 17 major depressive disorder (MDD) microarray studies. Impacts and effectiveness of the proposed inclusioin/exclusion evaluation on the final meta-analysis results were evaluated in the real examples. Additional simulations were performed to show the robustness and effectiveness of the proposed method. To our knowledge, this is the first systematic and objective quality assessment tool developed to decide inclusion/exclusion criteria for genomic meta-analysis. The QC measures and evaluation are described specifically for microarray meta-analysis in this paper but potentially can be generalized to genome-wide association studies (GWAS) or increasingly popular deep sequencing data sets.

2.2 METHODS

To assess the information quality of a microarray data set, we have sought several numerical measures from different perspectives. Overall we propose six quantitative quality control scores to find a cluster of studies which can be characterized as homogeneous, consistent, highly influential, and accurate for a further meta-analysis. Table 2 shows the summary of each quality control criterion.

Before presenting the specific quality control criteria, we first address the general structure of given data sets. Suppose we are trying to combine the K number of microarray studies, each study E_k $(k = 1...K)$ is denoted as

$$E_k = \{x_{kgs}\}_{1 \le g \le G;\, 1 \le s \le S_k} \quad \text{and} \quad \{y_{ks}\}_{1 \le s \le S_k}$$

, where $\{x_{kgs}\}$ represents the expression intensity of gene g and sample s in the study k, and $\{y_{ks}\}$ denotes the clinical outcome of sample s in the study k, which can be binary, multi-class, continuous, or censored; G represents the total number of genes when genes are

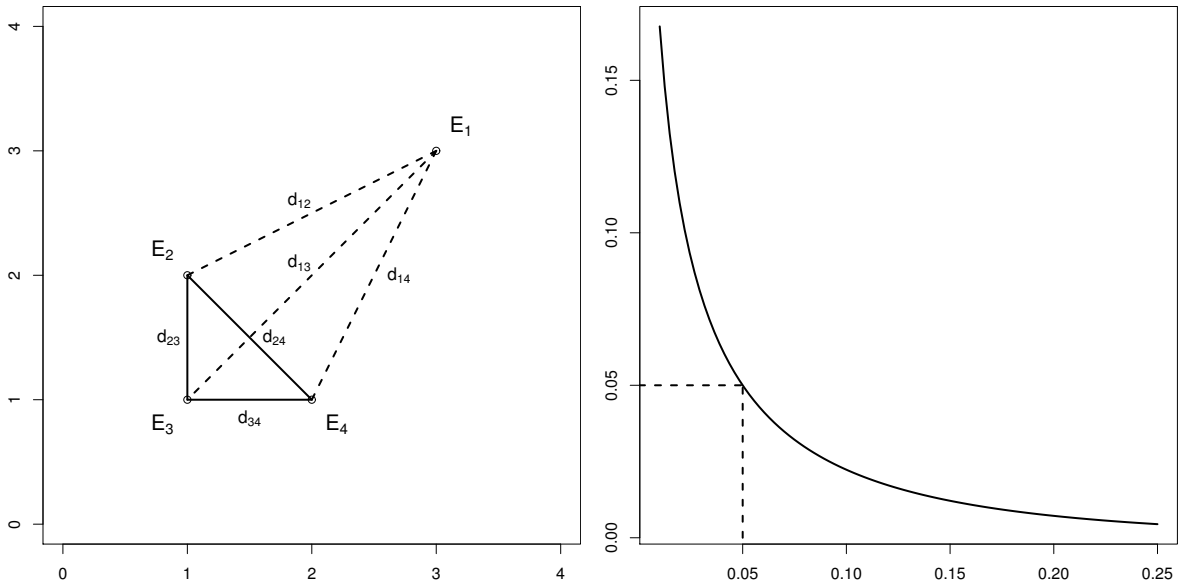
matched across studies, and S_k represents the number of samples in the study k. Also we define a g^{th} gene vector in the k^{th} study as $x_{kg} = (x_{kg1}, x_{kg2}, ..., x_{kgS_k})$.

2.2.1 Internal quality control (IQC) index

In this first criterion, the internal homogeneity of co-expression structure among studies is evaluated as an internal quality control (IQC) index. The IQC is a comparative measure that compares pair-wise differences among studies in an unsupervised manner (without any prior or external information other than the expression profile data) and the aim is to identify potential inconsistent or outlier studies from the quantified co-expression dissimilarity. We apply a concept of the correlation of correlations that was previously reported in the context of reproducibility analysis of gene co-expression patterns across studies, named as integrative correlation coefficients [28, 77]. Consider K studies to be combined. For a given study k, we define $\rho_{kij} = cor(x_{ki}, x_{kj})$ as the Pearson correlation coefficient of gene expression intensities between gene i and gene j in study k. The similarity between two studies m and n is defined as $r_{mn} = spcor\left((\rho_{mij};\ 1 \le i \le j \le G_{mn}), (\rho_{nij};\ 1 \le i \le j \le G_{mn})\right)$, which is the Spearman's rank correlation of the pairwise correlation structure between study m and n. The dissimilarity (or distance) between study m and n is defined as $d_{mn} = (1 - r_{mn})/2$. For a given study k, consider the set of distances from all other studies to the study k (i.e. $\tilde{D}_k^* = \{d_{kn}\}_{1 \le n \le K; n \ne K}$) and the set of all pairwise distance that do not involve study k (i.e. $\tilde{D}_k^{\#} = \{d_{mn}\}_{1 \le m \ne n \le K; m \ne K, n \ne K}$). When study k is an outlying study that contains co-expression structure very different from all other studies, the distances in $\tilde{D}_k^*$ are generally much greater than those in $\tilde{D}_k^{\#}$. Consider the two sets of distances follow certain probability distributions: $\tilde{D}_k^* \sim \mathcal{F}_1$ and $\tilde{D}_k^{\#} \sim \mathcal{F}_2$. We perform a hypothesis testing based on $H_0 : \mathcal{F}_1 = \mathcal{F}_2$ vs. $H_a : \mathcal{F}_1 = \mathcal{F}_2$ and apply one-sided Wilcoxon rank-sum (a.k.a. Mann-Whitney U) test [65] to generate a p-value, $P_{IQC}(k)$. Figure 1a shows an example that study 1 has a very different co-expression structure from other three studies and we will compare (d_{12}, d_{13}, d_{14}) and (d_{23}, d_{24}, d_{34}) by Wilcoxon rank-sum test to obtain a small p-value for study 1.

The hypothesis setting gives small p-value when study k is an outlying study. We apply a reverse transformation g(p) such that large p-value correspond to an outlying study. The

transformation will make the minus-log transformation below a quality score and consistent to the other QC measures to be introduced later. To keep the 0.05 statistical significance threshold invariant in the transformation, we define g as $g(p) = 1 - \mathcal{F}_{D_2}(\mathcal{F}_{D_1}^{-1}(p))$, where $D_1 \sim \mathcal{N}(z_{.95}, 1)$ and $D_2 \sim \mathcal{N}(-z_{.95}, 1)$. For example, g(0.05)=0.05, g(0.5)=0.0005 and g(0.01)=0.17. Figure 1b shows a plot of the transformation. Finally, the IQC measure of study k is defined as $IQC(k) = -log_{10}\, g\,(P_{IQC}(k))$. We will use log base 10 for all QC measures in this paper. Large IQC indicates that the study has homogeneous co-expression structure with other studies and is considered good quality to be included for meta-analysis.



(a) Concept of IQC (b) Inverse Transformation of P-value

Figure 1: **Internal Quality Control.** (a) Three points in the lower left represent homogeneous studies, and a point in the upper right is a heterogeneous study which has larger pair-wise distance to others. (b) X and Y axis are p-values before and after applying transformation g; as the result, smaller p-values mean stronger homogeneity.

2.2.2 External quality control (EQC) index

Compared to the unsupervised approach achieved in IQC, we develop a supervised quantitative criterion, named as external quality control (EQC) criterion. External knowledge of

pathways (i.e. functional or co-regulated gene sets) obtained from established databases (e.g. KEGG, GO, Biocarta and MSigDB) is applied to evaluate its consistency with a given study and subsequently to determine the study quality. We use similar pairwise gene correlation structure concept used in IQC and define an association measure between study k and a given pathway (gene set) w by

$$\begin{aligned} t_k &= t_k\left(\{\rho_{kij}\}_{1\le i,j\le G_k}; w\right) \\ &= \left(\frac{\sum_{j>i;i,j\in w}|\rho_{kij}|^l}{|w|\cdot(|w|-1)/2}\right)^{1/l} \Bigg/ \left(\frac{\sum_{1\le i<j\le G_k}|\rho_{kij}|^l}{|G_k|\cdot(|G_k|-1)/2}\right)^{1/l} \end{aligned}$$

, where ρ_{kij} is the Pearson correlation of gene i and gene j in study k as defined in IQC, the numerator is the l-norm average of absolute pairwise correlation in pathway w, the denominator is the corresponding l-norm average in the background genome G_k, and $|w|$ and $|G_k|$ are the number of genes in the pathway w and study k. If pathway w is relevant to disease status or experimental perturbation, we expect that the l-norm average among the pathway in the numerator will be much larger than that among genome background in the denominator and t_k will be significantly greater than 1. In this association measure, we disregard the sign of correlation coefficients and use l-norm to inflate differential impact of high and low correlations in the measure. We will use $l = 2$ throughout this paper to down-weight medium to low correlation coefficients and gives higher relative weight to large correlation coefficients (e.g. 0.82=0.64 and 0.32=0.09). We set up hypothesis testing $H_0 : t_k = 1 \quad \text{vs.} \quad H_a : t_k > 1$ and apply Monte-Carlo permutation analysis to obtain the empirical null distribution of the test statistic t_k [10, 71]. Specifically, we randomly sample from G_k a random pathway $w^{(b)}$ of equal size (i.e. $|w^{(b)}| = |w|$) in the bth simulation, calculate the corresponding $t_k^{(b)}$ and repeat for B times (b=1,...,B). The resulting p-value of the test is calculated as $P_{EQC}(k; w) = (\sum_{b=1}^{B} I(t_k^{(b)} > t_k) + 1)/(B + 1)$, where $I(\cdot)$ is an indicator function. Here, we adopt a conservative procedure to add 1 to both denominator and numerator in p-value calculation, considering the observed statistics is one of the simulated cases [71]. The EQC measure is then defined as $EQC(k; w) = -log_{10}\, P_{EQC}(k; w)$. Similar to IQC, large EQC(k;w) indicates that the study has significantly higher association with pathway

w in terms of gene pairwise correlation structure and is thus considered good quality to be included for meta-analysis.

We further extend the EQC measure for multiple pathways. Consider among M pathways $W = \{w_m, 1 \leq m \leq M\}$ available, a significant portion of them have high association measure with study k. We define a Fisher's score [25] by $S_k = -2 \sum_{m=1}^{M} logP_{EQC}(k; w_m)$. If the pathways are independent, the S score follows a chi-squared distribution with degree of freedom 2M. However, since the biological pathways always have hierarchical structure and high overlapping, we perform permutation analysis for B times to obtain simulated $S_k^{(b)}$. The resulting p-values is calculated as $P_{EQC}(k; w) = (\sum_{b=1}^{B} I(S_k^{(b)} > S_k) + 1)/(B + 1)$, and the EQC measure is similarly defined: $EQC(k; W) = -log_{10} P_{EQC}(k; W)$.

Comparing IQC and EQC, we note that EQC relies on a good selection of pathway set W and the evaluation of one study is independent from other studies. IQC, on the other hand, is a relative measure that depends on other studies under consideration but does not require external biological information.

2.2.3 Accuracy quality control (AQCg and AQCp) and consistency quality control (CQCg and CQCp) index

In the third and fourth criteria, we propose an accuracy quality control (AQC) and a consistency quality control (CQC) criteria that are aimed to quantify the reproducibility (accuracy or consistency) of differentially expressed genes (or pathways) detected in an individual study compared to those detected by meta-analysis. For AQCg for study k, the identified DE gene list from meta-analysis excluding study k (using Fisher's method under FDR=r%) is served as a gold standard. The DE gene list detected by study k (using Student's t-test with Benjamini-Hochberg procedure under FDR=r%) is then compared to the gold standard to generate a 2×2 table and calculate the sensitivity, specificity and Youden's index [121] (defined as sensitivity+specificity-1). One-sided Fisher's exact test can be used to determine the association (reproducibility) of DE gene list identified by meta-analysis and that identified by study k (H_0: the two gene lists have no association. vs. H_a: the two gene lists have

association). The p-value for study k is calculated from hypergeometric distribution:

$$P_{AQCg}(k) = \sum_{t=t_k}^{min(T^(k),T^{(-k)})} \frac{\binom{T^(k)}{t}\binom{G_k - T^(k)}{T^{(-k)} - t}}{\binom{G_k}{T^{(-k)}}}$$

, where G is the total number of genes, $T^(k)$ is the number of DE genes detected by study k, $T^{(-k)}$ is the number of DE genes detected by meta-analysis excluding study k and t_k is the number of DE genes detected both by study k and by meta-analysis excluding study k (see the 2×2 table in Table 1). The AQCg score is defined as $AQCg(k;r) = -log\ P_{AQCg}(k;r)$. We use FDR threshold $r = 5$ but can relax it to 10 or 20 when the data have weak signal. Large AQCg measure for a given study k indicates that DE genes produced by study k is reproducible compared to DE genes detected by meta-analysis excluding study k. We extend AQCg to AQCp where DE genes in the AQCg definition are replaced by enriched pathways. The pathway enrichment can be obtained by simple Fisher's exact test under certain DE gene threshold or other gene set analysis methods (e.g. GSEA [103] or GSA [19]). In this paper, we used simple Kolmogorov-Smirnov test under FDR=5% threshold to obtain enriched pathways.

In contrast to evaluating DE gene lists from a hard threshold in AQCg, we also apply an alternative of consistency quality control (CQC) measure by evaluating the consistency of differential expression ranking from single study analysis and meta-analysis. Specifically, ranks of differential expression evidence of study k is first calculated by Student's t-test and defined as $R_g^{(k)}$ for gene g and study k. From meta-analysis (using Fisher's method) excluding study k, the ranks of differential expression evidences are denoted as $R_g^{(-k)}$. The Spearman rank correlation between two rank vectors is defined as $\rho_k = spcor\left((R_g^{(k)}; 1 \le g \le G_k), (R_g^{(-k)}; 1 \le g \le G_k)\right) = 1 - 6 \cdot \frac{\sum_{g=1}^{G_k}(R_g^{(k)} - R_g^{(-k)})^2}{G_k(G_k^2-1)}$. To test $H_0 : \rho_k = 0$ vs. $H_a : \rho_k > 0$, we can approximate that $t = \rho_k \cdot \sqrt{\frac{G_k-2}{1-\rho_k^2}}$ follows a Student's t distribution with $G_k - 2$ degree of freedom under null hypothesis [52]. The resulting p-value is calculated as $P_{CQCg}(k) = 1 - \mathcal{F}_{G_k-2}(\rho_k \cdot \sqrt{\frac{G_k-2}{1-\rho_k^2}})$, where $\mathcal{F}_{G_k-2}$ represents the cumulative distribution function (cdf) of Student's t-distribution with $G_k - 2$ degree of freedom. The CQCg score is defined as $CQCg(k) = -log_{10}P_{CQCg}(k)$. Having a large CQCg measure for a given study k indicates that DE evidence produced by study k is relatively consistent with

DE evidence generated by meta-analysis excluding study k. We can also extend CQCg to CQCp where DE evidence and gene ranking in the CQCg definition are replaced by enriched pathways.

Table 1: Contingency table for AQC inference

	TRUE	FALSE	TRUTH
TRUE	t_k	$T^{(-k)} - t_k$	$T^{(-k)}$
FALSE	$T^k - t_k$	$F^{(-k)} - T^k + t_k$	$F^{(-k)}$
Observed	T^k	F^k	G_k or W_k

Table 2: Summary of Six Quality Control Scores

Types	Evaluation Criteria	External Pathway Knowledge Needed?	Clinical Outcome Needed?
IQC	Homogeneity of co-expression	No	No
EQC	Consistency of co-expression	Yes	No
CQCg	Consistency of gene ranking	No	Yes
CQCp	Consistency of pathway ranking	Yes	Yes
AQCg	Accuracy of detected biomarkers	No	Yes
AQCp	Accuracy of detected pathways	Yes	Yes

2.2.4 Visualizatioin and summarization for decision

We apply principal component analysis biplots [46] to assist the visualization and decision for inclusion or exclusion of studies in meta-analysis. Each microarray study is projected from high dimensional QC measures to a two dimensional PC subspace. The direction of each quality control measure is juxtaposed on top of the two-dimensional subspace using arrows. Specifically, the coordinates of each quality criterion are determined by its correlation to two

driven PCs. The origin of the biplot is taken as the statistical threshold with Bonferroni correction (i.e. projected from $log_{10}(0.05/\#studies)$ in each of the QC measure dimensions), suggesting that studies located in the opposite area of arrows are candidate outlier studies. The scale of each QC measure is standardized before PCA to avoid dominance of a particular QC measure due to scale problem. In addition to biplot visualization, we also define a quantitative summary score by calculating the ranks of each QC measure among all studies and then compute the standardized mean rank (SMR) of each study: (mean rank of all QC meausres)/(# of studies). By definition, $0 < SMR \leq 1$. Since CQCg and AQCg are usually highly correlated in our examples and CQCp and CQCp are highly correlated, we combine each pair of them by average into new QC measures, named as consistency and accuracy quality control (CAQCg and CAQCp) measures.

Note that our visualization and summarization tools are not meant for an automated recommendation. In the examples we explored, there are three categories in QC results: bad quality for definite exclusion, good quality for definite inclusion and borderline cases. Definite exclusion cases are often on the opposite side of arrows in the PCA biplots and have small QC ranks (and thus small SMR score). These studies are strongly suggested to be excluded from meta-analysis. On the other hand, definite inclusion cases are on the same side of arrows in the PCA biplots and have large rank scores. They are clearly good quality studies that should be included. Borderline studies happen to be in between the two extreme cases. Although an automated quantitative decision is desirable, it is often not practical. One should seek additionial qualitative assessment for the causes of bad quality, no matter for definite bad studies or borderline studies.

2.2.5 Application, evaluation and simulation in real datasets

We have evaluated our proposed method to four examples: brain cancer (7 studies), prostate cancer (9 studies), Idiopathic Pulmonary Fibrosis (IPF) (8 studies), and major depressive disorder (MDD) (17 studies). Summary of these studies is listed in tables 3 - 6. The most microarray data sets were collected from public repositories such as NCBI Gene Expression Omnibus [18] and EBI ArrayExpress [76], or web pages directed in the original papers.

Several non published data sets were obtained from the labs of Dr. Kaminski and Dr. Sibille. Most data sets are already normalized by original authors. When raw data are available, RMA [45] was applied for preprocessing. To obtain a robust result, we have applied a gene filtering in each study level, which removes 40% of non-expressed genes based on the expression intensity and 40% of non-informative genes based on variance. Gene matching across studies is done by matching official gene symbols using Bioconductor [29] packages. When multiple probes match to one gene symbol, the probeset which has the largest IQR is selected.

In EQC criterion, external pathways are needed for calculating EQC measures. We only considered pathways that have at least 5 genes in each study. Conceptually, using pathways relevant to the disease or experimental perturbation will generate better EQC evaluation. For cancer studies, we chose to use GSEA Biocarta v3.0 pathways [103] since the pathways are cancer specific. A total of 217 Biocarta pathways were used in the brain cancer example. For prostate cancer studies, the overall data quality and information seemed to be weaker and we chose only the top 50 pathways among the 217 pathways for better performance (top pathways were identified by combined p-values using Fisher's method). For MDD studies, 99 pathways were selected from all GSEA MSigDB v3.0 by the keyword search using a list of relevant terms provided by Dr. Sibille: GABA, INSULIN, DIABETES, IMMUNE, THYROID, ESTROGEN, DEPRESSION, AGING, ALZHEIMERS, PARKINSONS, and HUNTINGTONS. For IPF studies, we have chosen top 50 pathways out of all 6769 number of GSEA MSigDB v3.0 pathways in terms of the EQC measure. For AQCp and CQCp measures, pathway database is also needed to generate enriched pathways before evaluation. We use all MSigDB c2 v3.0 pathways for both AQCp and CQCp in all four examples.

We performed 100,000 simulations in the permutation analysis of Fisher scores in EQC measure and thus the largest range of EQC measure is limited to 5 (that corresponds to p=1E-5). For AQC measures, we applied two-sample Student's t-test and Kolmogorov-Smirnov test for AQCg and AQCp, respectively. All p-values were adjusted by Benjamini-Hochberg procedure [3] to control FDR at the level of 0.05. Fisher's method (sum of minus log-transformed p-values) was used for meta-analysis in both AQC and CQC evaluation when performingn meta-analysis of all studies except for the study k. In AQC measures,

weak signal examples may generate only few DE genes or pathways that makes the AQC measure invalid or unstable. We chose a liberal cutoff (unadjusted $p < 0.05$) to prevent each lower score from having zero value because of weak signal.

To assess the validaty and performance of our proposed method, we performed downstream analysis to assess its impact on DE gene and pathway detection and performed simulation to assess the accuracy of detecting problematic studies. The results are reported in the following section. All implementation was written by R statistical language [83]. An R package, "MetaQC" is publicly available online at CRAN (http://cran.r-project.org/)

2.3 RESULTS

2.3.1 Quality assessment in four examples

Table 3 - 6 lists summary information of studies in the four examples. For brain cancer example, we have obtained 7 brain cancer studies comparing Anaplastic Astrocytoma (AA) and Glioblastoma multiforme (GBM) samples (Table 3). Dreyfuss et al. [17] have combined four studies for meta-analysis of which three are used in this analysis. Figure 2A shows the PCA biplot of the result and Table 7 shows the detailed QC measures and SMR score. The first two PCs in Figure 2A explains about 92% of total variance, and all scores are highly positively correlated with the first PC. The scores marked with asterisks in Table 7 indicate non-statistical significance ($p > 0.05/\#$ of studies), which means the specified study might have an adverse effect on meta-analysis based on the QC measure. The Yamanaka study is clearly below statistical threshold and has low values in all measures; it is a definite exclusion case that should be excluded from the meta-analysis. On the other hand, the top 5 studies in Table 7 performed very well for all criteria, indicating that they are definite inclusion cases for meta-analysis. The Paugh study (study 6 in Figure 2A), is however a borderline case. The QC measures were mostly low and just passed the statistical significance. Interestingly, when scrutinizing the causes of poor quality of Yamanaka and Paugh studies, Yamanaka uses a different platform and both studies are of smaller sample size.

In the second example, we applied proposed QC assessment to 9 prostate cancer studies comparing normal and primary cancer patients. The data details are summarized in Table 4. QC results are shown in Figure 2B and Table 8. Compared to brain cancer studies, we found that prostate cancer studies were mostly performed in earlier years with older array platforms. Although the first two PC also capture high percentage of variance (93%), the studies more scattered around and even good studies had quite different performance by different QC criteria. For example, Varambally and Wallace had better score on IQC and EQC but not CQC and AQC while Welsh, Lapointe, and Singh, had better performance in CQC and AQC but not IQC and EQC. Yu had performed the best in all criteria. In considering sample size, array platform and QC measures, we suggest to exclude the bottom 3 studies: Nanni, Tomlins, and Dhanasekaran from meta-analysis and mark Singh as a borderline case. The worse performance of prostate cancer studies shown here reflects the fact that prostate cancer is a heterogeneous cancer [89].

As a third example, we evaluated 8 Idiopathic Pulmonary Fibrosis (IPF) studies which identify signature genes of IPF patients compared to normal. IPF is one of the most lethal chronic lung disease, and its mean survival is only 3-5 years regardless of treatment [54]. Table 5 shows data summary, Figure 2C demonstrates the PCA biplot and Table 9 lists the details of QC scores. Interestingly, although these 8 data sets are mostly from very different microarray platforms, at least five of them performed very well, indicating good quality for meta-analysis. Of the three worst QC studies, Emblom utilized a cDNA array platform which might caused the worst performance. Yuga and Larsson both have small sample size (n=7 for Yuga and n=12 for Larsson) which might be the reasons of low QC scores. The two top studies, KangA and KangB, are unpublished data from Dr. Kaminskis lab with large well-characterized cohorts.

In our final example, we apply QC evaluation to 17 Major Depressive Disorder (MDD) studies that compare normal and MDD patients. These 17 studies are obtained from post-mortem brain tissues of various brain regions and are considered very weak signal, small sample size and heterogeneous data sets. The details of each data set is in Table 6. The QC results are shown in Figure 2D and Table 10. In Figure 2D, noticeably many studies scattered near the origin because of weak signal of most studies. From Table 10, the top

3-5 studies are clear definite inclusion studies and the bottom five studies are definite exclusion studies. Other studies are borderline cases somewhat in the middle. Most CQCg and AQCg scores are significantly lower than other examples since each individual MDD study is weak signal and the DE gene and pathway detections are relatively unstable. We note that the 7 out 9 bottom studies were all from Stanley Foundation Tissue Bank, which has been suspected to have worse quality from problematic tissue collection and processing.

Table 3: 7 Brain Cancer Studies (AA vs. GBM)

Arthor	Year	Platform	Sample Size	Source
Freije[27]	2004	HG-U133A,B	85	GSE4412
Phillips[80]	2006	HG-U133A,B	100	GSE4271
Sun[104]	2006	HG-U133 Plus 2	100	GSE4290
Yamanaka[118]	2006	Agilent	29	GSE4381
Petalidis[79]	2008	HG-U133A	58	GSE1993
Gravendeel[34]	2009	HG-U133 Plus 2	175	GSE16011
Paugh[78]	2010	HG-U133 Plus 2	42	GSE19578

Table 4: 9 Prostate Cancer Studies (Normal vs. Primary)

Arthor	Year	Platform	Sample Size	Source
Dhanasekaran[15]	2001	cDNA	32	www.pathology.med.umich.edu
Welsh[112]	2001	HG-U95A	34	public.gnf.org/cancer/prostate/
Singh[96]	2002	HG-U95Av2	102	www.broad.mit.edu/
Lapointe[58]	2004	cDNA	103	GSE3933
Yu[122]	2004	HG-U95Av2	146	GSE6919
Varambally[109]	2005	HG-U133 Plus 2	13	GSE3325
Nanni[69]	2006	HG-U133A	30	GSE3868
Tomlins[107]	2006	cDNA	57	GSE6099
Wallace[111]	2008	HG-U133A2	89	GSE6956

Table 5: 8 IPF Studies (Normal vs. IPF)

Arthor	Year	Platform	Sample Size	Source
Pardo[75]	2005	Codelink	24	GSE2052
Yang[119]	2007	Agilent 43K	29	GSE5774
Larsson[59]	2008	HG-U133 Plus 2	12	GSE11196
Vuga[110]	2009	Codelink	7	GSE10921
Konishi[55]	2009	Agilent 4x44K	38	GSE10667
Emblom[22]	2010	cDNA	58	GSE17978
KangA	2011	Agilent 4x44K	63	-
KangB	2011	Agilent 8x60K	96	-

Table 6: 17 MDD Studies (Normal vs. MDD)

Data Name	Year	Platform	Sample Size	Source
MD1_AMY	2009	HG-U133 Plus 2	28	Dr. Sibille
MD3_AMY	2009	HumanHT-12	42	Dr. Sibille
MD1_ACC	2009	HG-U133 Plus 2	32	Dr. Sibille
MD3_ACC	2009	HumanHT-12	44	Dr. Sibille
MD2_ACC_M	2010	HG-U133 Plus 2	18	Dr. Sibille
MD2_ACC_F	2010	HG-U133 Plus 2	26	Dr. Sibille
MD2_DLPFC_M	2010	HG-U133 Plus 2	28	Dr. Sibille
MD2_DLPFC_F	2010	HG-U133 Plus 2	32	Dr. Sibille
NY_DLPFC_M	2004	HG-U133A	26	Dr. Sibille
NY_oFC_M	2004	HG-U133A	24	Dr. Sibille
Feinberg	-	HG-U95Av2	27	www.stanleygenomics.org
KatoB	2004	HG-U95Av2	26	www.stanleygenomics.org
Kemether	-	HG-U133p	24	www.stanleygenomics.org
AlartC	-	HG-U133A	22	www.stanleygenomics.org
SklarA	-	HG-U95Av2	23	www.stanleygenomics.org
SklarB	-	HG-U95Av2	23	www.stanleygenomics.org
Sokolov	-	HG-U95A	26	www.stanleygenomics.org

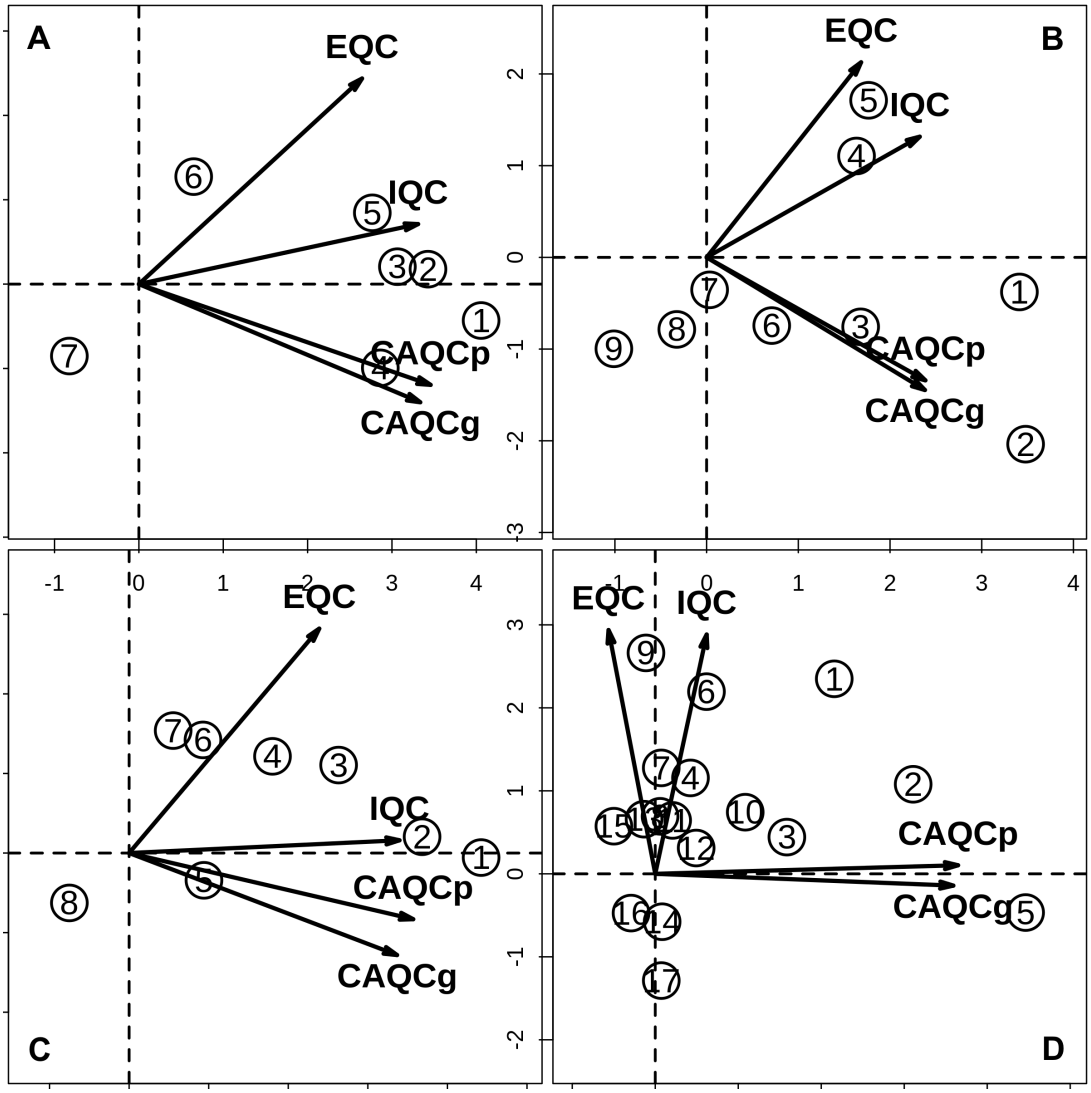


Figure 2: **Integrative quality control.** Each circled number represents the overall rank of a study. Smaller number corresponds to higher quality study. A) 7 brain cancer studies: the 7th ranked studies, Yamanaka, looks definite outlier study; the 6th study, Paugh, is in the borderline; the top five studies look solid. B) 9 prostate cancer studies: the top five studies look good, the 6th ranked study is in the borderline, and the bottom three studies, Nanni, Tomlins, and Dhanasekaran, are outliers. C) 8 IPF studies: the top four studies looks good; the three studies ranked between 5 to 7 are in the borderline; the 8th ranked study, Emblom, is a definite outlier. D) 17 MDD studies: the bottom four studies look outliers; the top ten looks fine, and between 11th and 13th ranked studies look to be in borderline.

Table 7: Brain Cancer Study - Quality Control Scores

#	Study	IQC	EQC	CQCg	CQCp	AQCg	AQCp	Rank
1	Sun	4.96	5.00	307.65	251.33	152.83	108.37	1.88
2	Freije	5.42	5.00	239.31	158.73	118.06	81.62	2.62
3	Phillips	4.52	5.00	242.36	146.71	106.59	69.19	3.62
4	Petalidis	4.11	3.16	274.25	171.48	111.27	101.10	3.75
5	Gravendeel	6.64	4.70	98.37	107.06	47.67	63.89	4.00
6	Paugh	1.51*	5.00	5.00	3.60	2.31	9.84	5.12
7	Yamanaka	0.10*	0.78*	1.69*	2.26	1.58*	0.71*	7.00

Table 8: Prostate Cancer Study - Quality Control Scores

#	Study	IQC	EQC	CQCg	CQCp	AQCg	AQCp	Rank
1	Yu	7.74	3.23	52.43	64.14	19.95	38.30	2.25
2	Welsh	5.04	2.12*	68.59	101.31	26.46	54.66	2.50
3	Lapointe	4.06	2.28	26.36	59.42	7.00	33.90	3.75
4	Varambally	4.68	4.70	15.38	21.15	4.18	13.21	3.88
5	Wallace	7.95	4.22	0.00*	28.70	0.00*	2.13*	4.75
6	Singh	2.14*	2.05*	19.60	28.74	4.61	24.17	5.25
7	Nanni	1.92*	1.92*	2.22*	6.01	2.00*	13.61	6.75
8	Tomlins	2.67	0.52*	3.76	3.65	1.19*	6.12	7.38
9	Dhanasekaran	0.01*	0.63*	0.01*	0.23*	0.04*	0.10*	8.50

Table 9: IPF Study - Quality Control Scores

#	Study	IQC	EQC	CQCg	CQCp	AQCg	AQCp	Rank
1	KangA	6.64	5.00	307.65	146.87	96.71	90.88	1.88
2	KangB	5.57	5.00	273.67	114.30	84.37	69.74	2.62
3	Konishi	6.89	5.00	58.19	42.70	25.50	57.20	2.75
4	Yang	4.34	5.00	41.70	56.35	14.20	29.43	3.88
5	Pardo	4.07	2.08*	25.14	38.84	20.60	25.05	5.38
6	Vuga	2.28	5.00	1.37*	26.25	1.77*	18.01	5.38
7	Larsson	1.79*	5.00	0.59*	1.88*	0.52*	3.21	6.25
8	Emblom	0.03*	1.12*	0.83*	0.57*	0.43*	1.98*	7.88

Table 10: Major Depressive Disorder Study - Quality Control Scores

#	Data Name	IQC	EQC	CQCg	CQCp	AQCg	AQCp	Rank
1	MD2_ACC_F[1]	9.80	5.00	19.22	40.01	6.17	27.47	3.75
2	MD2_DLPFC_M[1]	3.22	5.00	56.70	41.02	9.37	33.27	4.38
3	MD2_ACC_M[1]	3.76	3.05	24.59	33.78	3.80	17.16	6.62
4	MD1_ACC[1]	3.41	5.00	10.59	19.28	0.39*	10.21	6.62
5	MD2_DLPFC_F[1]	3.05	1.12*	52.05	62.94	14.32	46.79	7.00
6	NY_oFC_M[1]	11.56	3.74	0.12*	18.09	0.4*	13.32	7.25
7	NY_DLPFC_M[1]	4.05	5.00	1.63*	14.82	0.3*	6.61	7.88
8	MD3_AMY[1]	0.96*	5.00	3.23	12.03	1.54*	7.05	8.50
9	KatoB[2]	11.54	5.00	0*	1.46*	0.45*	2*	8.62
10	Kemether[2]	8.01	1.91*	12.21	8.92	9.79	1.63*	8.75
11	MD3_ACC[1]	1.37*	4.70	8.70	15.65	1.8*	4.06	9.38
12	MD1_AMY[1]	3.09	2.97	1.49*	17.14	0.39*	16.76	9.62
13	SklarB[2]	0.73*	5.00	0*	9.71	0*	8.80	10.75
14	Sokolov[2]	4.07	0.3*	0.46*	1.4*	0.6*	6.85	10.75
15	Feinberg[2]	0.35*	5.00	0.32*	2.41*	0.17*	0.77*	13.12
16	SklarA[2]	1.2*	1.93*	0*	1.01*	0*	2.41*	14.75
17	AltarC[2]	0.69*	0.08*	0*	15.95	0*	0.91*	15.25

[1] Data from our collaborator, Dr. Etienne Sibille.
[2] Data from Stanley Foundation, suspected worse quality in the tissue collection and processing

2.3.2 Impacts on DE gene and pathway detection

To evaluate the ultimate biological impact of our MetaQC method, we investigated the marginal impact of a meta-analysis on DE gene and enriched pathway detection. We hypothesized that including an additional informative study in a meta-analysis would provide increased statistical power to detect more DE genes and enriched pathways. Figure 3A and 4A show the number of DE genes and enriched pathways detected under FDR=0.5%, respectively, when 7 brain cancer studies were added sequentially in the meta-analyses in the order of SMR score of MetaQC. Interestingly, the number of detected DE genes and pathways dropped clearly when including the two suspect problematics studies: Paugh and Yamanaka. The result supported the recommendation provided by MetaQC. This simple incremental analysis also argues the necessity of adequate inclusion/exclusion criteria in meta-analysis.

The results for prostate cancer (Figure 3B and 4B) and IPF examples (Figure 3C and 4C) demonstrated more complex situation than in brain cancer. The number of DE genes under FDR=0.1% generally increased as more studies were added while the number of detected pathways decreased when the 5th and the 6th studies were added in prostate cancer and IPF, respectively. In Supplement Figure 2B, we found that Wallace, Singh and Tomlins generally have stronger DE evidence than other studies. The increased number of detected DE genes in Figure 3B might have been caused by this bias although the pathway result in Figure 4B did not show increased finding. The prostate cancer example demonstrated a case that pure AQCg or CQCg method focusing on commonality of DE gene detection is not effective enough when studies are highly heterogeneous. In the IPF example, similar observation can be found. Inclusion of Emblom greatly increased the number of DE genes (Figure 3C) but decreased the number of detected pathways (Figure 4C). This may also be due to the larger number of DE genes detected by Emblom (Supplement Figure 2C).

Figure 3D and 4D shows the result of the MDD examples. In contrast to previous examples, MDD studies have very weak overall signals, so we applied very liberal DE gene detection criterion which is unadjusted p-value=1%. However, in terms of pathway identification, we could get a similar number of enriched pathways with usual FDR=5% threshold. In spite of its liberal threshold, the number of DE genes is smaller than other examples.

Also the number of increased DE genes as more studies included is less than others except Kemether and AltarC which made DE genes increased significantly than its previous step. However, as we notice from previous two examples, these two studies are not considered as quality studies because their inclusion caused significant drop in the number of identified pathways in the figure 4D. Again, supplement figure 2D shows the large number of DE genes in both studies compared to others.

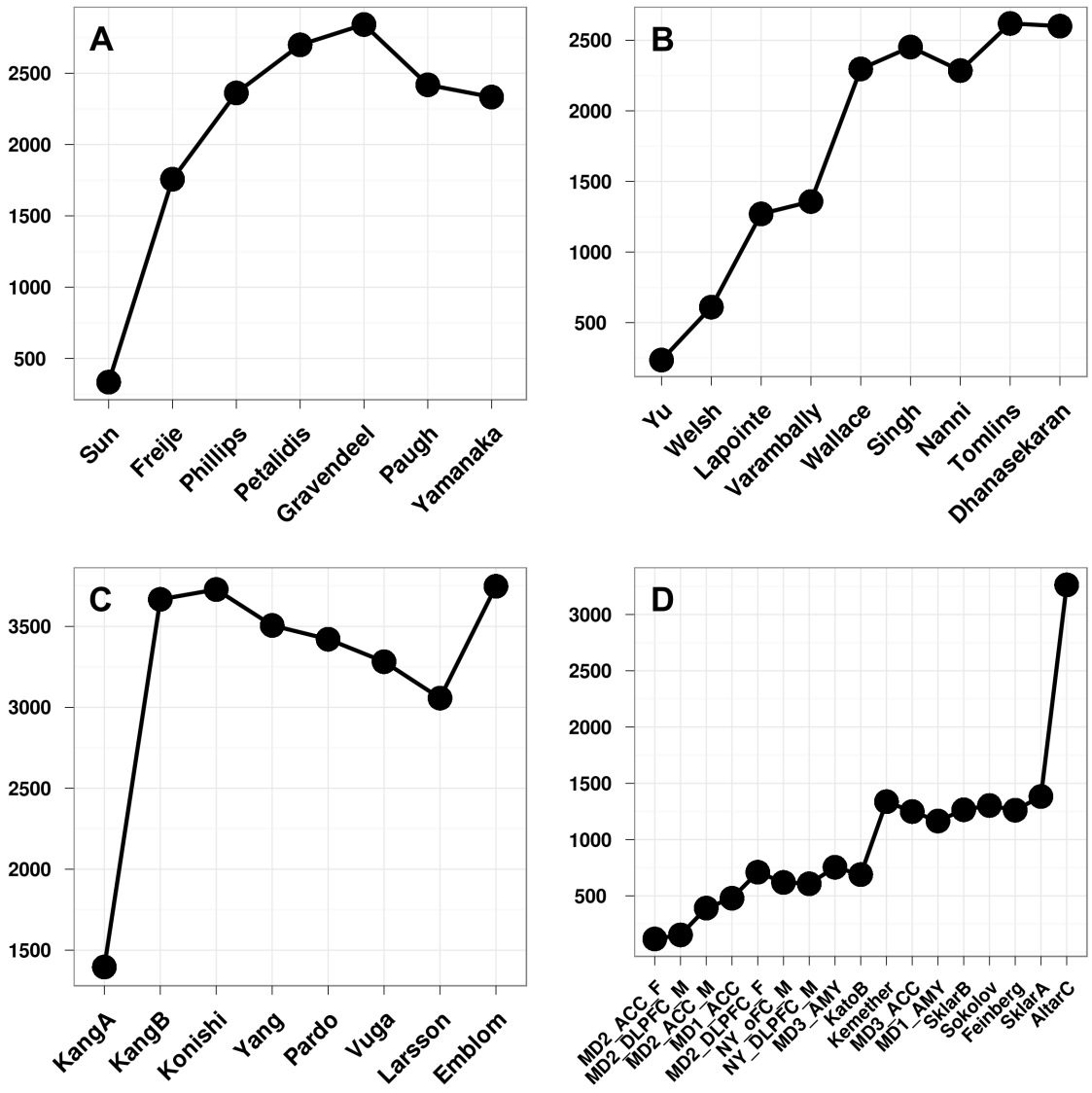


Figure 3: **Marginal impacts on DE genes detection.** X-axis represents each study included cumulatively to a series of meta-analyses. The order of addition follows the quality rank in the table 7 - 10. Y-axis represents the number of DE genes. A) The example of 7 brain cancer studies to investigate the marginal impact of including one additional study. FDR 0.005 was used for DE genes detection. B) The example of 9 prostate cancer studies. The number of DE genes in y-axis are detected with FDR 0.001. C) The example of 8 IPF cancer studies. FDR 0.001 was used for DE genes detection. D) The example of 17 MDD studies. The number of DE genes in y-axis are detected with liberal p-value 0.01.

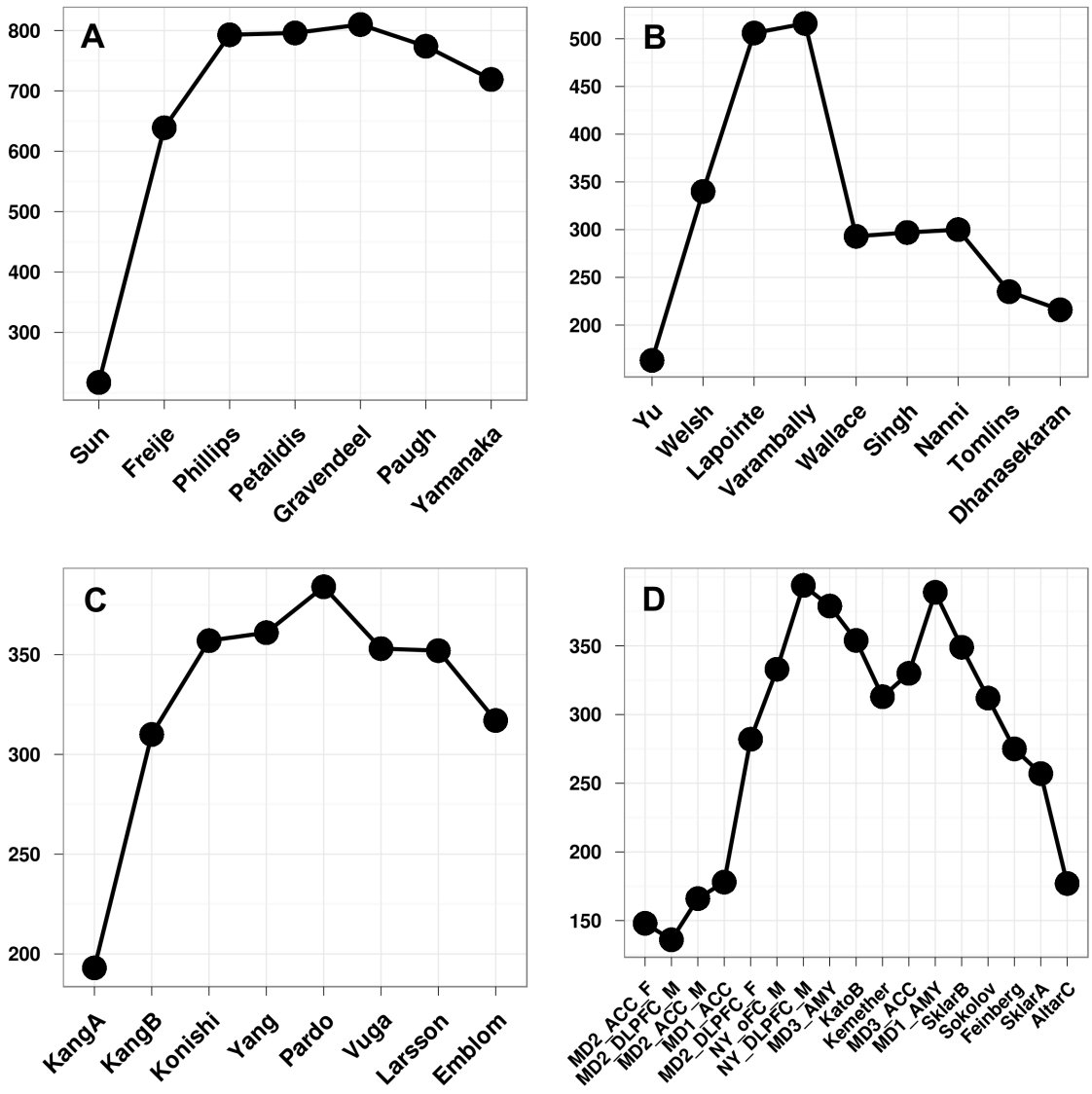


Figure 4: **Marginal impacts on enriched pathways detection.** X-axis represents each study included cumulatively to a series of meta-analyses. The order of addition follows the quality rank in the table 7 - 10. Y-axis represents the number of enriched pathways identified by FDR 0.05. A) The example of 7 brain cancer studies to investigate the marginal impact of including one additional study. B) The example of 9 prostate cancer studies. C) The example of 8 IPF cancer studies. D) The example of 17 MDD studies.

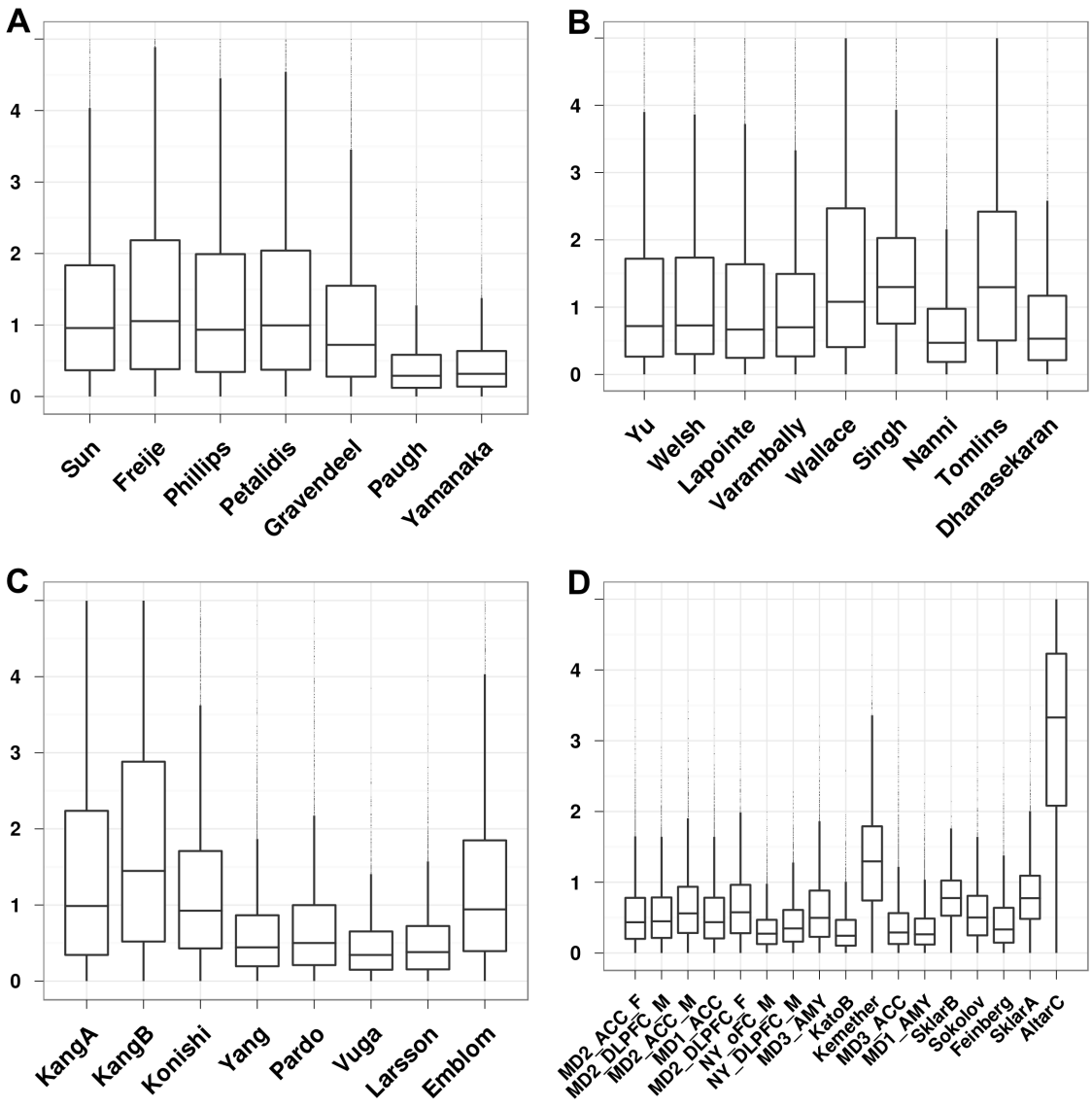


Figure 5: **P-value distribution of each example studies.** X-axis represents studies which are ordered by the quality rank in the table 7 - 10. Y-axis represents -log10 transformed p-value of each gene. A) In 5 brain cancer studies, top 5 studies have comparable distribution of p-values, and the worst two studies had relatively weak signals. B) In 9 prostate cancer studies, the overall proportion of significant p-value in Wallace, Singh, and Tomlins was higher than others. C) In 8 IPF studies, the overall proportion of significant p-value in Emblom was higher than others, though it had the worst quality. D) In 17 MDD studies, the overall proportion of significant p-value in Kemether and AltarC was much higher than others.

2.3.3 Simulations

To further validate the QC result of our proposed method, we investigated a simple yet insightful simulation scheme. For example, seven brain cancer studies were fixed for MetaQC evaluation. In each simulation, an additional prostate cancer study is added as a known outlier. The simulations were repeated through all prostate cancer studies and the changes of SMR scores were recorded and compared. In Figure 5A, the 1-SMR scores of seven brain cancer studies were plotted in the first columns (labelled as "NA"). In the following nine simulations, a prostate cancer studies was added to the seven brain cancer studies and the 1-SMR scores were recalculated. The added outlier study was plotted by an asterisk symbol. The result shows that the added prostate cancer studies always generated small 1-SMR score similar to Yamanaka study and were always detected as a definite exclusion case. The result also demonstrated the effectiveness and robustness of our proposed method to perform QC assessment and screen out outlier studies. Moreover, the addition of a random irrelevant study as the "null" study provides a more objective and practical threshold to decide the exclusion of studies. In this context, the decision in brain studies seems evident that Yamanaka should be excluded. For the second simulation in Figure 5B, a brain cancer was added as an outlier study to nine prostate cancer studies in each simulation. The results showed that the added brain cancer study had 1-SMR scores better than Nanni, Tomlins and Dhanasekaran. Since brain cancer and prostate cancer share some commonalities as two types of cancers, the added brain cancer studies can serve as a baseline negative control to argue exclusion of the three bottom studies. In Figure 5C, we further added brain cancer studies as outliers into the 8 IPF studies. The result showed similar pattern that argues to exclude Embolm and Larsson studies. Figure 5D shows the result that one of 7 brain studies are added to 17 MDD studies sequentially. It looks the top 7 studies have the best quality, and the bottom 5 studies have the worst quality, and the middle 5 studies are in the borderline.

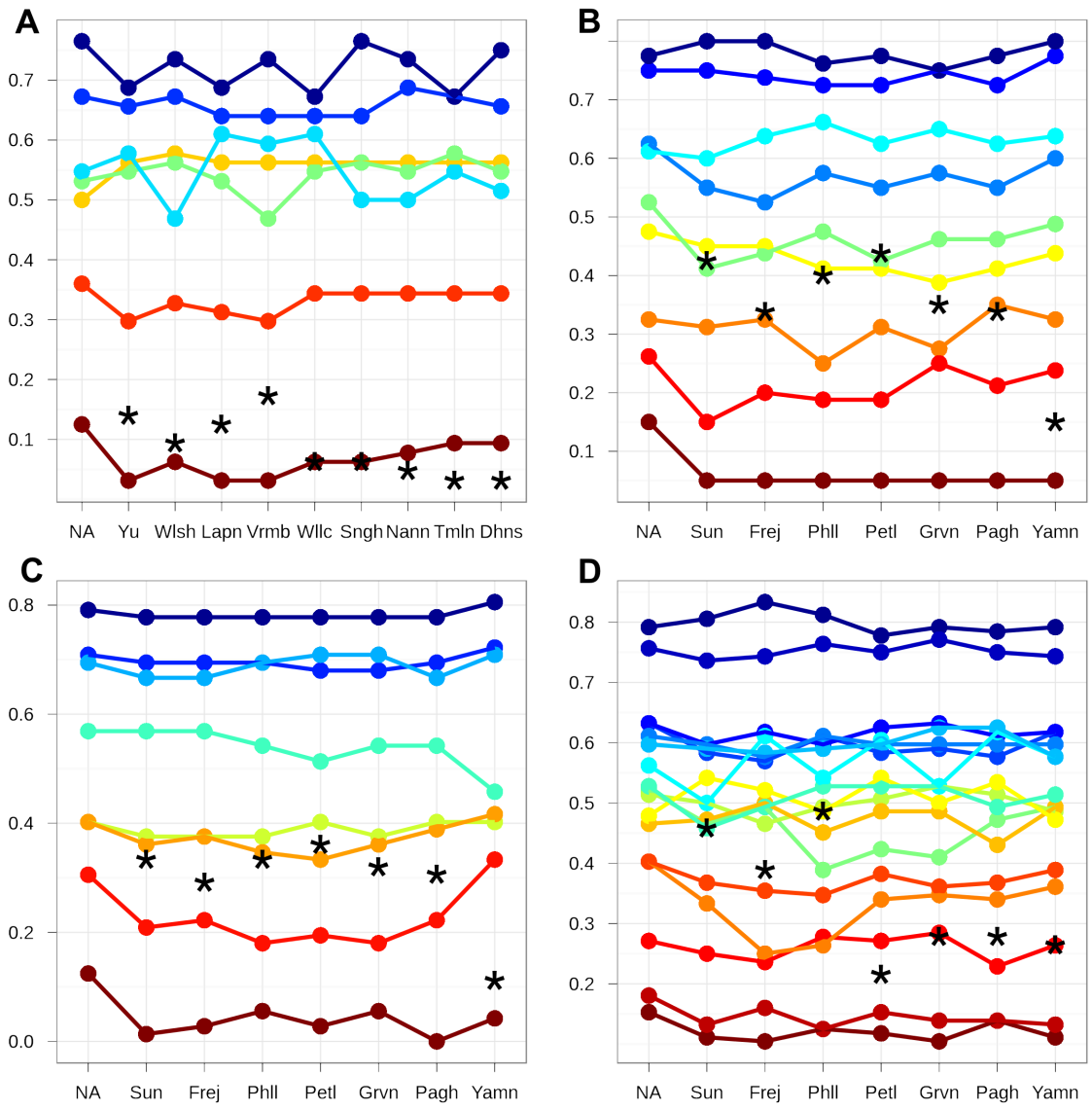


Figure 6: **Simulation study showing the effect of adding an irrelevant study.** Y-axis represents 1 - standardized mean rank (SMR); high quality studies should have greater values. X-axis represents the addition of a totally irrelevant study to one of current four examples. A set of irrelevant studies were obtained from one of the other three examples. The order of inclusion was based on the quality rank in the table 7 - 10, so high quality irrelevant studies have smaller x values. NA represents the original quality result without an irrelevant study which is the same result as table 7 - 10. A black asterisk represents the 1-SMR of the added irrelevant study. Studies under the asterisk should be outliers. A) 7 brain cancer studies. B) 9 prostate cancer studies. C) 8 IPF studies. D) 17 MDD studies.

2.4 CONCLUSIONS AND DISCUSSIONS

We have proposed 6 quantitative inclusion/exclusion criteria which measures the quality of studies. The goal of our methods is to find a group of homogeneous studies for the better result in a meta-analysis. As we can see in the example of prostate cancer studies, many of them are very heterogeneous. Especially, the striking opposite relationship between the sharp increase in the number of DE genes in the figure 3 and the sharp decrease in the number of enriched pathways in the figure 4 suggests that the all inclusive study selection strategy may lead a very different conclusion in the same type of meta-analysis. In contrast, as we can find in the simulation results, our proposed method incorporated various aspects of quality assessment criteria so that it can screen out outlier studies in an effective and robust way and eventually lead to consistent results between meta-analyses.

Although our approaches are not aimed to give a fully automated decision threshold, we showed that a PCA visualization tool and a simple simulation procedure could be served as complementary tools to find a reasonable decision boundary without considering any qualitative information. When it is possible to obtain any reasonable domain knowledge, the qualitative information would be very useful in the case that when the decision boundary is very complex as we saw in the MDD example.

Lastly, we observed the importance of homogeneity of studies in a meta-analysis in various analysis including DE gene detection and enriched pathway identification. Especially, our newly developed MetaPCA methods showed the significant increase of prediction accuracy using the top 5 brain cancer studies; however, the effect in top 5 prostate studies was not significant as the brain studies suggesting that the amount of homogeneity among considered studies are the barometer of the success of a meta-analysis.

3.0 METAPCA : META-ANALYSIS IN THE DIMENSION REDUCTION OF GENOMIC DATA

(This paper is in preparation to submit to Annals of Applied Statistics)

Principal Component Analysis (PCA) enables us to explore high dimensional genomic data through projection to a low-dimensional space. As an exploratory tool to visualize subjects in 2 or 3 dimensional subspace while minimizing information loss, PCA is one of the most popular multivariate analysis techniques. In this paper, we consider simultaneous dimension reduction using PCA when multiple studies are combined. Although similar concepts of common principal components analysis exist, the advantage of such a practice in the meta-analysis context has not been studied. We propose two basic ideas to find a common PC subspace by eigenvalue maximization approach and angle minimization approach, and we extend the concept to incorporate Robust PCA and Sparse PCA in the meta-analysis framework. We evaluated the advantages and limitations of our methodology in the context of dimension reduction for data visualization and supervised machine learning using five examples of real microarray data, we show the information gain obtained by adopting our proposed procedure. We also suggest that for the successful meta-analysis, homogeneous data selection procedure is necessary.

3.1 INTRODUCTION

The Principal Component Analysis (PCA) is one of the most popular techniques that enable us to explore high dimensional data through projection to a low-dimensional space. Each Principal Component (PC) is orderly derived to be a linear combination of original features

in that it explains as much of the variance (information) as possible. The first PC is chosen to have the largest variance, and the next PCs are sequentially derived to have the largest variance in the orthogonal subspace of selected PCs [46].

One of the many applications of PCA is to visualize high dimensional subjects in two or three dimensional PC subspace. The derived subspace is the optimal choice to represent as much information (variance) as possible in such a reduced dimension, i.e. it minimizes the sum of squares of the projection errors. Another possible application of PCA is to utilize PCs as an input for other statistical methodologies such as regression analysis or various multivariate techniques [46, 39, 53]. Two main advantages of PC based approaches are to overcome the multicollinearity problem which occurs when highly correlated variables are included together in an analysis and to alleviate the curse of dimensionality which numerous multivariate methods are suffered from [4, 2].

In many cases, PCA has been very successful and gives a concise presentation of overall data structure. Particularly, when PCA is applied to microarray data, it is hoped that the chosen PCs elucidate the informative low dimensional manifold such as interesting patterns or clusters of subject or genes [85, 92, 63, 11, 120]. However, microarray data are often noisy, and PCA is sensitive to outliers; often times it fails to get an effective representation in the reduced dimension. Another issue of PCA is its difficulty of interpretation when PC is a linear combination of a large number of variables.

Robust PCA and Sparse PCA have been pursued in the literature to overcome these shortcomings. To achieve robustness of PCA, various Robust PCA approaches were proposed: influence function techniques [41, 42, 43, 14], multivariate trimming [31], alternating minimization [51], and random sampling techniques [24]. Recently, [8] proposed a low-rank component recovery based approach, which is reportedly to attain strong performance gain compared to previous methods; and [12] proposed an appealing algorithm using projection pursuit such that a robust measure of variance is maximized in lower dimensional space.

On the other hand, to achieve better interpretation of each PC, Sparse PCA has recently gained increasing popularity: a maximal variance approach [47], a regression framework [123], greedy search and exact methods using branch-and-bound techniques [67], a convex relaxation/semidefinite programming approach [13], a regularized low-rank matrix approximation

approach [93], penalized matrix decomposition based approach [115], and a generalized power method [48]. Particularly, the method proposed by Witten and Tibshirani [115] was shown to unify the approaches of Shen [93], Jolliffe [47], and Zou [123]. The main advantage of Sparse PCA is to find sparse loadings so that each PC can be interpretable by a smaller set of original features, which is essential in microarray analysis where thousands of features often exist in the data. It also can be served as a good technique for an unsupervised feature selection.

In this paper, we consider the situation that we have several similar studies and try to compare several PCA results in parallel. By traditional PCA, each PC subspace is not comparable. A naïve solution for this problem is to project new data sets to previously derived PC subspace. However, this approach is not efficient in that it fails to utilize all information from available data sets and not robust in that it depends totally on the quality of target PC subspace. Better approaches should be the ones that can use all information and are robust such that the driven PCs retain only informative common source of variance and reveal the true separation among subjects.

Here, we propose two approaches that resolve the issues occurred by an individual PCA. We hypothesize that several PCA results from compatible data sets have an advantage to elucidate the common cause of variance throughout data sets regardless of individual noise and heterogeneity of each study. Focusing on the commonness among studies is the philosophy of what we are trying to do with MetaPCA by finding the "optimal common subspace" in multiple studies.

Although some similar ideas have been in the literature, in our knowledge there was no effort to apply the concept of MetaPCA to genomic researches. Current literatures regarding to genomic meta analysis are wholly focused on biomarker detection or gene set enrichment analyses. Here, we have investigated the simultaneous dimension reduction of multiple microarray studies using MetaPCA, and we show the usefulness of common subspace in terms of dimension reduction for data visualization and classification.

The structure for the rest of the paper is as follows: In section 3.2, we propose two MetaPCA optimization criteria (eigenvalue maximization and angle minimization approaches) to find the common subspace. In section 3.3, we extend the preferred angle minimization

approach to Robust PCA and Sparse PCA scenarios. In section 3.4, we apply the proposed methods to five examples of microarray meta-analysis. The results are evaluated quantitatively under data visualization and supervised machine learning framework. We provide discussions and conclusions in section 5.

3.2 METAPCA

3.2.1 Common PC Subspace

The idea of comparing and combining subspaces generated by Principal Components (PC) was first studied by Krzanowski [56]. The eigenvectors of several groups of individuals has been compared by computing the angles between the subspaces spanned by the first k PCs of each group. In addition, Krzanowski has proposed a simple estimate of the common subspace as eigenvectors of the sum of sample covariance matrices and used it for the likelihood ratio test of common principal components [57, 26]. This paper is based on these ideas of finding common subspaces of several data sets; however, while those methods are focused on checking and testing the similarities of covariance matrices of each data set, we are more interested in developing integrative methods to represent the information gain by utilizing several relevant data sets simultaneously in one analysis.

It is interesting to find informative linear combinations of features which can represent multiple data sets simultaneously. This goal was described as a hypothesis for the test of Common Principle Components (CPC) as follows [26, 57] :

$$H_b : L^t \Omega_i L = \Lambda_i \, (i = 1, ..., K)$$

, where L is an orthogonal (p $\times$ p) matrix and the Λ_i is an diagonal matrix, and Ω_i is the sample covariance matrix of i^{th} data set—there are total K number of studies considered. Although this hypothesis is only the case when the number of features (p) is smaller than the number of subjects (n), it can be easily extensible to the problems of when p is greater than n. In that case, L should be an orthogonal matrix (p $\times$ n*), where the n* is the minimum number of sample sizes among considered data sets.

Here we alternatively define a meta-subspace as the common subspace of considered data sets which can fulfill specific optimization criteria which are described in the next. And we propose two basic methods to find meta-subspaces: the first one is aimed to find the best subspaces which maximize the sum of variance of each data set, named as Eigenvalue Maximization Approach; the second one is intended to find the best subspaces that minimize the sum of squared cosine angle between meta-subspace and each individual subspace, named as Angle Minimization Approach.

3.2.2 Eigenvalue Maximization Approach

The first approach to find a meta-subspace can be described as an optimization problem of

$$\begin{aligned} \max \sum_{i=1}^{K} \Lambda_i &= \underset{L}{\operatorname{argmax}} \sum L^t \Omega_i L \\ &= \underset{L}{\operatorname{argmax}}\, L^t \left(\sum \Omega_i\right) L \end{aligned} \tag{3.1}$$

$$\max \sum_{i=1}^{K} w_i \frac{\Lambda_i}{\lambda_{i1}} = \underset{L}{\operatorname{argmax}}\, L^t \sum w_i \frac{\Omega_i}{\lambda_{i1}} L \tag{3.2}$$

, where λ_{i1} and w_i represent the largest eigenvalue and the weight of i^{th} data set, respectively. In the equation 3.2, the normalization of each covariance matrix was done by dividing with its largest eigenvalue, λ_{i1}; so the largest possible value of $\max \sum_{i=1}^{K} \frac{\Lambda_i}{\lambda_{i1}}$ is the same as the number of studies, K; additionally, if a weight of i^{th} study, w_i, is obtained from the external information, we can apply weighted sum of covariance matrices. The objective function 3.1 is a special form of 3.2 when all data sets have the equal weights and their scale difference of covariance matrices are ignored so that no normalization is applied. The solution of these optimization problem is the same as the one from the eigendecomposition of $\sum \Omega_i$ or $\sum w_i \frac{\Omega_i}{\lambda_{i1}}$.

3.2.3 Angle Minimization Approach

The second approach is based on the geometrical interpretation of PCs [33]. Suppose we have applied the K number of individual PCA analyses. Then we can get the K number of eigenvector matrices V_i which has each dimension of $p \times n_i$; n_i is less than or equal to the

rank of each data set. It was shown that a vector L in the original p-dimensional data space which minimizes the sum of all squared cosine angles between L and V_i can be calculated by eigendecomposition of $\sum V_i V_i^t$ [56]. The objective function can be denoted as follows:

$$\underset{L}{\operatorname{argmax}}\, L^t \left(\sum V_i V_i^t \right) L \tag{3.3}$$

$$\underset{L}{\operatorname{argmax}}\, L^t \left(\sum w_i V_i V_i^t \right) L \tag{3.4}$$

The objective function 3.3 is a special form of 3.4 when all data sets have the equal weights. The solution simply can be derived by the eigendecomposition, and the first k's of orthogonal vectors L are the loadings of the meta subspace.

3.3 EXTENSION OF METAPCA

Additionally we extended the basic MetaPCA idea to more sophisticated and recent advances of PCA, specifically Robust PCA and Sparse PCA.

3.3.1 Robust Angle Minimization Approach

Robust PCA was introduced as a generalization of PCA [61, 43]. Local optimal solution is found by Projection Pursuit [40], and a robust measure of variance (e.g. IQR, MAD, and Qn) is used as a projection index instead of variance. The advantage of Projection Pursuit approach is its fast calculation when only first few PCs are considered. Specifically, we can generalize objective function of PCA in terms of Projection Pursuit as follows:

$$v_k = \underset{\|v_k\|=1, v_k \perp v_1, \ldots, v_k \perp v_{k-1}}{\operatorname{argmax}} PI(v_k^t x_1, \ldots, v_k^t x_n) \tag{3.5}$$

, where v_k is the k^{th} eigenvector, and PI is called "projection index" and is the same as variance in the case of usual PCA. The idea of robust PCA is simple that robust measure of variance is used as projection index instead of variance.

We adopted GRID algorithm [12] to find robust PC directions, V_i^*, of each data set and simply replaced the objective function 3.3 as follows:

$$\underset{L}{\operatorname{argmax}}\, L^t \left(\sum V_i^* V_i^{*t} \right) L$$

The idea of GRID algorithm is to take advantage of the easy optimization of 3.5 in two dimensional space. In two dimension, the objective function 3.5 is reduced to optimize

$$\begin{aligned} \theta^* &= \underset{\theta; -\pi/2 \le \theta \le \pi/2}{\operatorname{argmax}} PI\left(cos\theta V_i + sin\theta e_j\right) \\ V_i^* &= cos\theta^* V_i + sin\theta^* e_j \end{aligned}$$

, where e_j is the canonical basis vector $(j = 1, ..., p)$.

3.3.2 Sparse Angle Minimization Approach

We extended the framework of [123] in the case of $p \gg n$ to incorporate multiple data sets in a single SPCA procedure. The original objective function for a single component can be written as follows:

$$\underset{\alpha,\beta}{\operatorname{argmin}} \|X - X\beta\alpha^T\|_F^2 + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1 \text{ subject to } \|\alpha\|_2 = 1$$

, where $\lambda_1, \lambda_2 \ge 0$, and β and α are p-vectors; $\|\cdot\|_F^2$ represents the squared Frobenius norm, which is the sum of squared elements of the matrix.

Consider S number of microarray data $(X_i; 1 \le i \le S)$, which each dimension is p number of features in the rows and n number of samples in the columns.

The Meta Sparse PCA (Meta-SPCA) algorithm is as follows:

1. For the calculation of each meta-eigenvector $L_j\,(1 \le j \le k)$, repeat from step 2 to 7
2. Let α_{ij} start at the first eigenvector of $X_i\,(1 \le i \le S)$
3. For fixed α_{ij},

$$\begin{aligned} \beta_{ij}^* &= \left(\sum_{i=1}^{S} \frac{|\alpha_{ij}^T X_i^T X_i|}{\|\alpha_{ij}^T X_i^T X_i\|} - \frac{\lambda_j}{2} \right)_+ Sign\left(\sum_{i=1}^{S} \frac{\alpha_{ij}^T X_i^T X_i}{\|\alpha_{ij}^T X_i^T X_i\|} \right) \\ \beta_{ij} &= \frac{\beta_{ij}^*}{\|\beta_{ij}^*\|} \end{aligned}$$

4. For fixed β_{ij}, compute SVD of $X_i^T X_i \beta_{ij}$, that is

$$X_i^T X_i \beta_{ij} = U_i D_i V_i^T, \text{then } \alpha_{ij} = U_i V_i^T$$

5. Repeat step 3 and 4 until convergence.
6. Apply Angle Minimization Method 3.3.1 to combine β_{ij} $(1 \leq i \leq S)$ as follows:

$$\underset{L_j}{\text{argmax}}\, L_j^t \left(\sum_{i=1}^{S} \beta_{ij} \beta_{ij}^T \right) L_j$$

7. Project X_i to the orthogonal space of L_j as follows:

$$X_i := X_i \left(I - L_j L_j^T \right)$$

One of advantages of Meta-SPCA is that it can tolerate missing genes among data sets since we can get zero values for the specified elements of α_{ij} from the augmented data set X_i^*, which has the augmented features which are the union of features among all considered data sets. We set the default value of λ as $\frac{\#\text{ of data sets}}{\sqrt{\#\text{ of augmented features}}}$ to consider each feature and component equally. Additionally, we allowed one to set the number of features for the analysis by deciding λ automatically.

3.4 APPLICATIONS

3.4.1 Data Description

3.4.1.1 Spellman data Spellman's yeast cell cycle data [99] was utilized to evaluate the performance of MetaPCA in terms of data visualization in two dimensional subspace. Biologically, this data set should be considered as four independent studies using four different synchronization methods: α arrest (alpha), arrest of cdc15 or cdc28 temperature-sensitive mutant (cdc15 and cdc28), and elutriation (elu). We filtered out genes which have overall missing values $\geq 10\%$ or log_2 transformed standard deviation $\geq .45$. 1025 genes were left, and the number of time points in the experiments were 18, 24, 17, and 14 for alpha, cdc15, cdc28, and elu, respectively. Additionally, we have imputed missing values using *impute.knn*

function in R statistical language [83]. In a recent report, the impact of missing value imputation to the downstream analysis such as classification was overall not severe [72]. We have utilized this data set for 2D dimension reduction for visualization.

3.4.1.2 Prostate cancer data We have used two sets of prostate cancer studies for two different applications: The first set of four studies have three classes of subjects: normal, primary, and metastasis [58, 107, 109, 122]. We used the first set for 2D dimension reduction for visualization. The second set of five studies have two classes of subjects: normal and primary [58, 109, 111, 112, 122]. We used the second set for the supervised learning (classification) evaluation. In both sets of data, we applied similar gene filtering and missing value imputation so that we have 3056 and 3016 genes in each set, respectively.

3.4.1.3 Mouse metabolism data This dataset involves samples from three genotype mice: wild-type (WT), LCAD knock-out (LCAD) and VLCAD knock-out (VLCAD). Deficiency of VLCAD is known to be related to a common energy metabolism disorder in children. On the other hand, LCAD-deficient mice are known to have impaired fatty acid oxidation and develop a disease similar to other disorders of mitochondrial fatty acid oxidation. For each of the 12 mice (four mice in each genotype), four types of tissues (brown fat, skeletal, liver and heart) were harvested and microarray experiments were performed to study the expression changes across genotypes. We applied similar gene filtering and missing value imputation so that we have 3175 genes left. We have used this data set for 2D dimension reduction for visualization.

3.4.1.4 Brain cancer data We also obtained five brain cancer studies which have both Anaplastic Astrocytoma (AA) and Glioblastoma multiforme (GBM) samples [104, 27, 79, 34, 80]. We applied similar gene filtering and missing value imputation so that we have 3004 genes left. The presumed goal of this meta-analysis is to find genomic difference between AA and GBM. We have used this data set to evaluate MetaPCA for classification.

3.4.2 Dimension Reduction For Data Visualization

The first application of simultaneous dimension reduction by MetaPCA is to find a common low (2D or 3D) dimensional subspace in which all data set can be comparable. As a classical example of visualization, we have considered Spellman yeast cell cycle data. The goal of the analysis is to show cyclic patterns of subjects which are correspondent to the elapsed time after initial cell cycle synchronization. Ideal figures are the ones which have two cyclic patterns in the first three experiments (alpha, cdc15, and cdc28) and one cyclic pattern in the last experiment (elu) from the known experimental conditions observed by microscope. Figure 7 shows the results from non-meta approach for a comparison purpose. Columns and rows represent each data and the subspace each data was projected, respectively. Diagonal plots are from usual PCA results, meaning each data was projected to its own subspace. Off-diagonal plots are projection of one study onto PC space generated by another study. For instance, the top rows are results when each of the four studies was projected to the alpha PC subspace. From the diagonal plots, we can tell that usual individual PCA may fail to capture the true data structure; instead, the truth can be revealed by borrowing information from the others. For example, cdc15 data in its own subspace had the worst result since it was hard to find the expected cyclic patterns—Previously, the oscillating nature of cdc15 synchronized data has been observed and investigated [62]. However, if cdc15 is projected to alpha space, we can observe clear cyclic patterns in the cdc15 data (See figure 7).

Direct projection approach—a study projected to a PC subspace of the other study—is a good alternative to borrow information from other data set; however, there are shortcomings that it does not perform real information integration. In the case of cdc15 example, since we knew what patterns to look for, we could tell alpha space was the best projection for cdc15 data. In general, such cyclic prior information is not available. It is reasonable to expect that automatic information integration of all four studies will identify a more informative PC subspace and provide better visualization.

Figure 8 shows results from the four proposed MetaPCA methods. Strikingly, we found clear cyclic patterns in almost every projection. Particularly, the improvement of cdc15 result was astounding. We could observe two clear cycles in alpha, cdc15, and cdc28 data,

and one cycle in elu data. Moreover, all four projections in each row represent the exact same projected space, allowing samples comparable between studies: we could even find that the numbers and cyclic change of direction are also comparable between plots. It is interesting that the Meta-SPCA results are based on only 40 genes (20 genes for each PC) and still represent the cyclic patterns quite well.

As a second example of dimension reduction for visualization, we considered four prostate cancer studies that have three types of samples: normal, primary, and metastasis. We expect the distribution of subjects in these three classes to have a transitional change: from normal to primary and then metastasis by disease severity. Figure 9 shows the result from angle minimization approach. We projected all four studies to the driven PC subspace by MetaPCA. As expected, we found a clear transitional pattern that the subjects were clustered by classes from normal to metastasis via primary. Interestingly, the most important separation among classes depends on the first meta PC, meaning that the first loading should have important information regarding to genes that separate those three classes. Moreover, the second PC separates well the metastasis from the others. From this example, we recognize that the first two PCs are essential in separating the three classes, but due to the nature of PCA, all genes have non-zero loading and the genes that are important for the separation were not clear.

Meta-SPCA focuses to make those PCs interpretable in terms of smaller number of non-zero loading genes. Figure 10 shows the result of Meta-SPCA using the same four prostate data as in Figure 9. We arbitrarily selected the number of non-zero loading genes to be 20 for a concise interpretation of each PC, and dimension reduction identifies a list of genes having strong relationship with prostate cancer. Genes with non-zero loading in the first two PCs are shown in the Table 11. Most of the top genes (19 out of the first component and 19 out of the second component; marked by asterisk) were known related to prostate cancer in PubMed and Google Scholar literature search. The advantage of Meta-SPCA is to find sets of informative genes in a sequential manner so that their order of finding is correspondent to the amount of overall information; in other words, the first set of genes which have non-zero coefficients in the first meta-PC loading are conceptually more informative genes, and the genes in the second loading are the next.

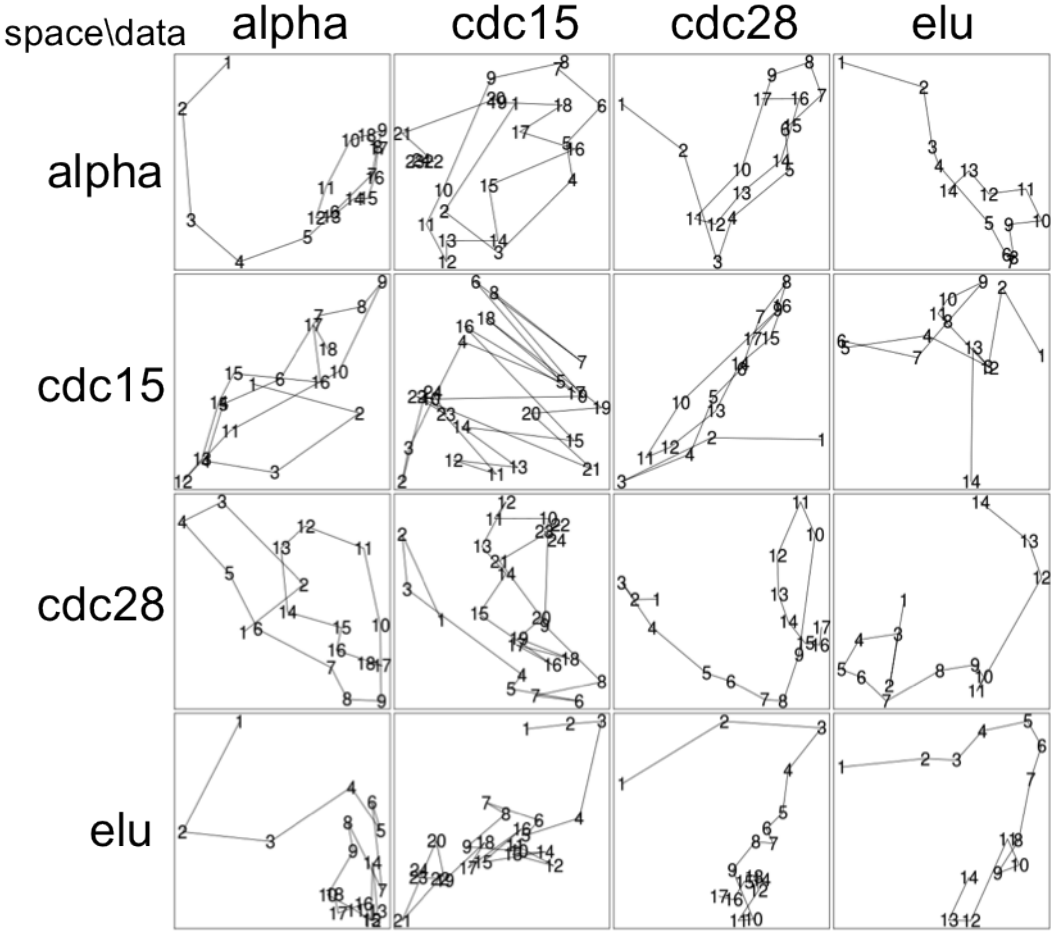


Figure 7: **PCA result for Spellman data.** Four data sets in Spellman yeast cellcycle data were projected to four PC subspace which is denoted as y-axis. The diagonal plots are when each data set is projected to its own PC subspace. The off-diagonal plots are when each data was projected to PC subspaces generated from other data. Although we can observe some cyclic patterns overall, some plots like cdc15 in the cdc15 subspace were hard to find cyclic patterns. Interestingly, cdc15 in the alpha subspace shows better cyclic patterns. From this result, cdc15 can be thought to be beneficiary for MetaPCA.

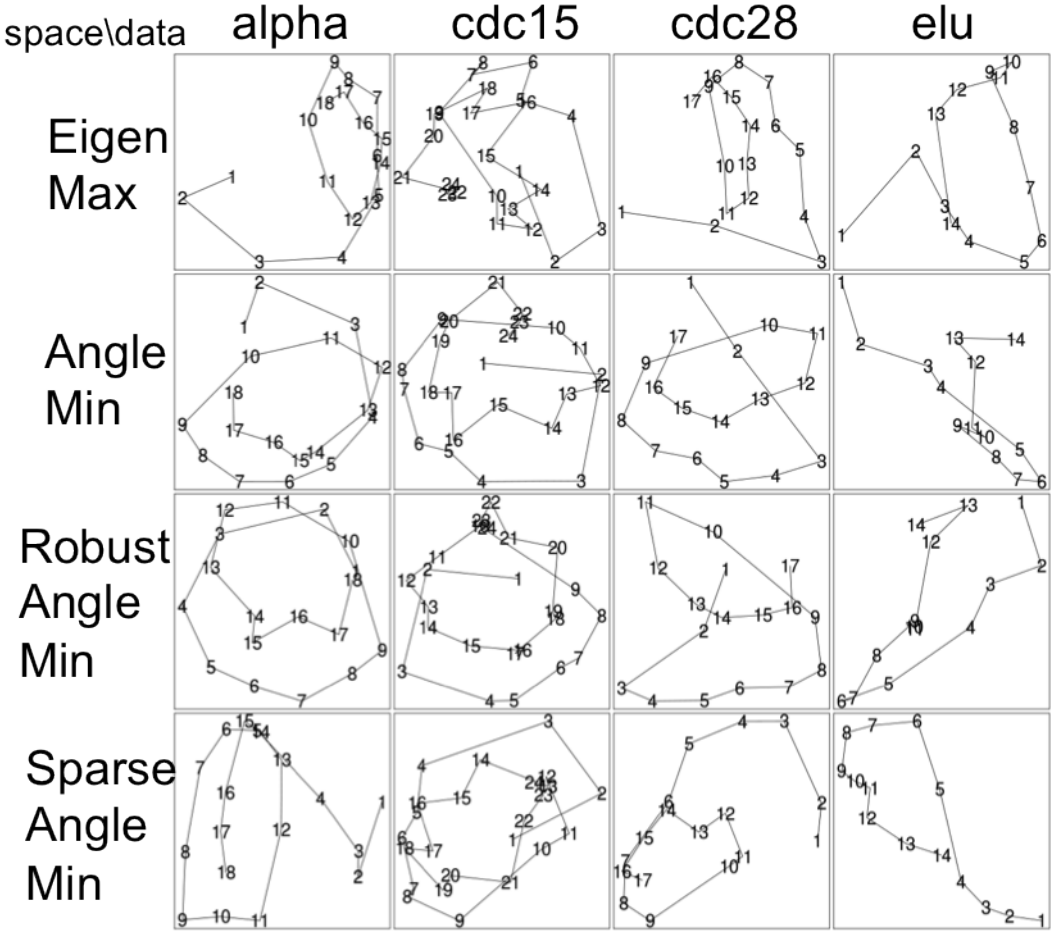


Figure 8: **MetaPCA result for Spellman data.** Four data sets in Spellman yeast cellcycle data were projected to four Meta-PC subspace which is denoted as y-axis. The first four subspace are from each data set. The diagonal figure is when each data set is projected to its own PC subspace. The first four off-diagonal figures are when each data was projected to PC subspaces generated from other data. The last two figures represent meta subspace found by eigen maximization and angle minimization approach, respectively. We can recognize the bottom two results show better cyclic patterns.

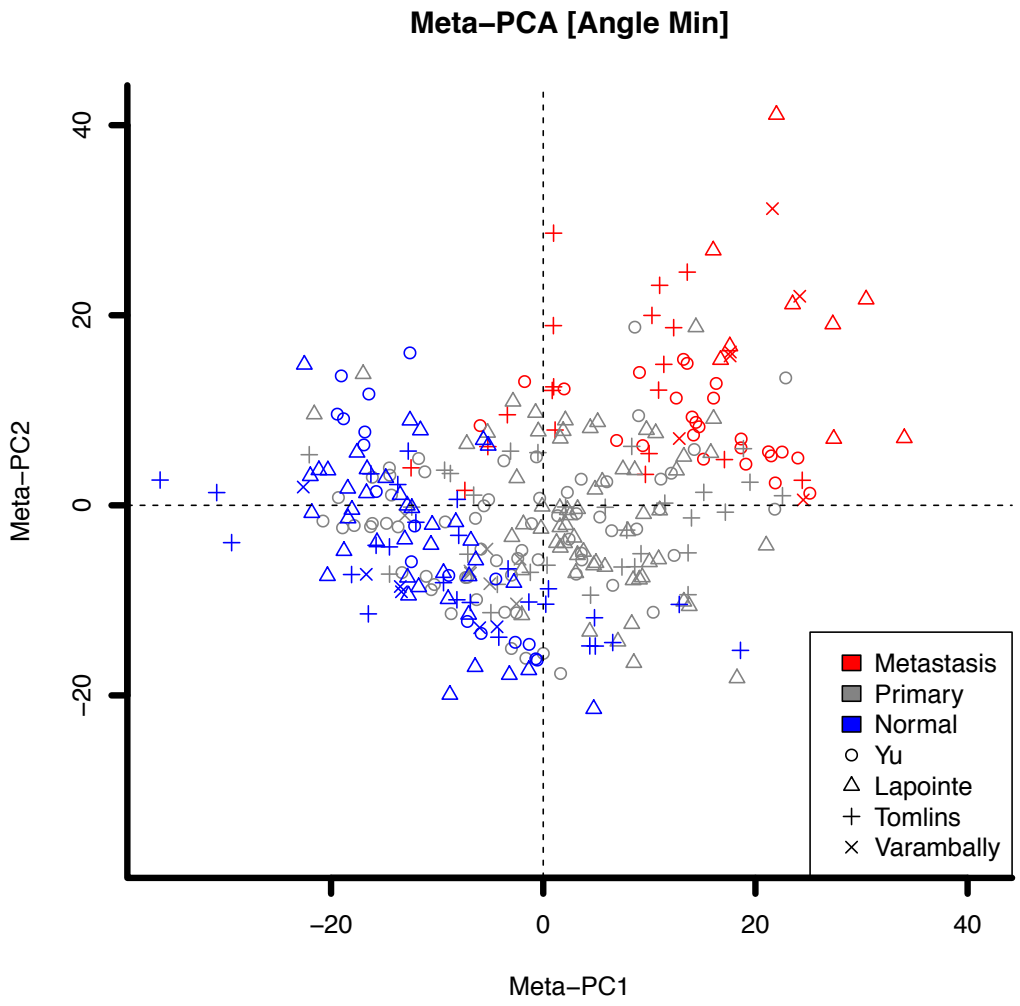


Figure 9: **MetaPCA result for Prostate cancer data.** Four prostate studies were projected to the 2D meta-subspace found by angle minimization method. Each data set has three disease classes, which is represented as different colors. Subjects from each study are represented as different shapes. Regardless of difference in data sets, the same classes tend to be clustered together. Moreover, it is for sure that there exists an order of distribution which follows the disease classes, i.e. from normal to metastasis via primary. The first component separates all three classes, and the second component separates metastasis from the others.

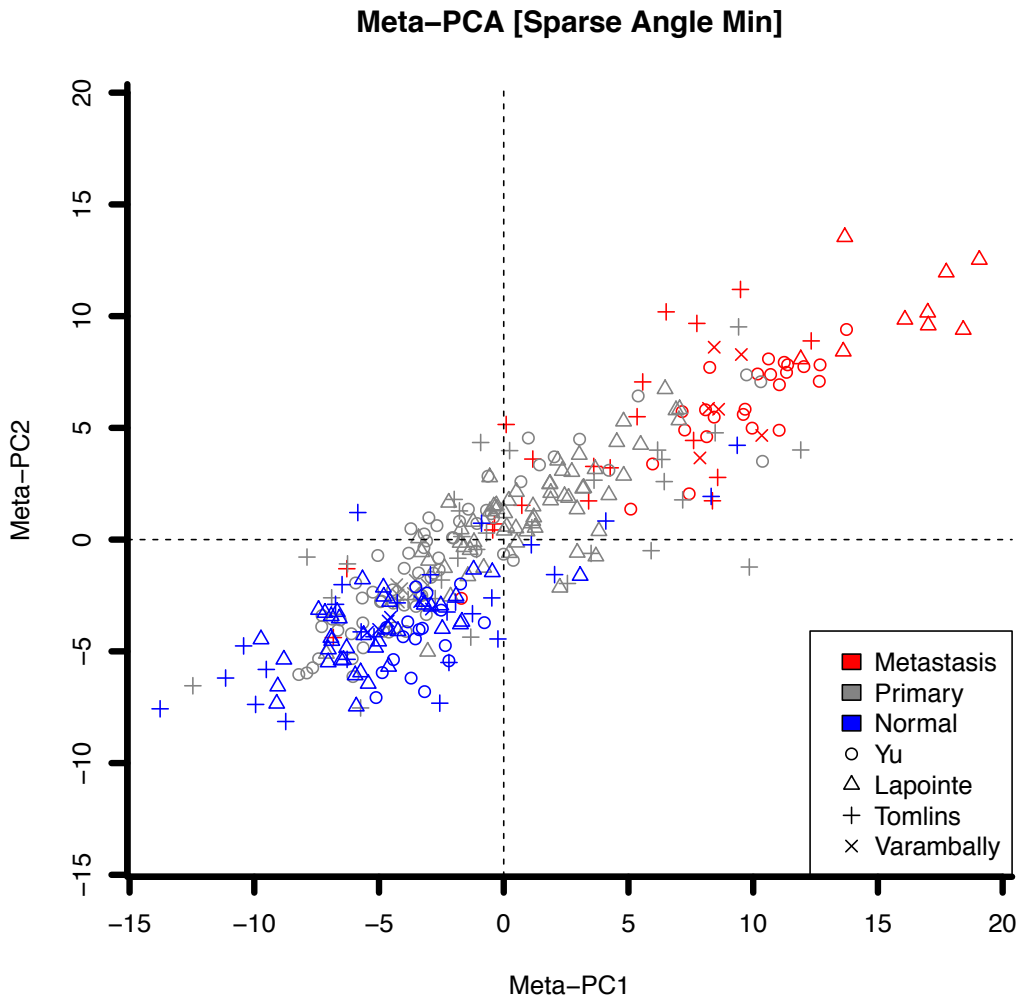


Figure 10: **MetaSPCA result for Prostate cancer data.** Four prostate studies were projected to the 2D meta-subspace found by Meta-SPCA. Each data set has three disease classes, which is represented as different colors. Subjects from each study are represented as different shapes. Regardless of difference in data sets, the same classes tend to be clustered together. Moreover, it is for sure that there exists an order of distribution which follows the disease classes, i.e. from normal to metastasis via primary.

Table 11: First Two Loadings of Meta-SPCA

Gene Symbols	PC1	PC2
ACTG2*	-0.549	0.000
SLC14A1*	-0.353	0.000
MYH11*	-0.316	0.000
LTF*	-0.297	0.000
CNN1*	-0.277	0.000
PLN*	-0.229	0.000
C10orf116*	-0.190	0.000
KRT5*	-0.169	0.000
PCP4*	-0.168	0.000
FOSB*	-0.166	0.000
MMP7*	-0.150	0.000
MAOB*	-0.136	0.000
FHL1*	-0.134	0.000
LMOD1*	-0.127	0.000
MFAP4	-0.119	0.000
ACTA2*	-0.113	0.000
PTGS2*	-0.108	0.000
CTGF*	-0.099	0.000
FBLN1*	-0.089	0.000
CYR61*	-0.078	0.000
EDNRA*	0.000	-0.321
IGF1*	0.000	-0.314
TAGLN*	0.000	-0.311
RARRES1*	0.000	-0.309
CAV1*	0.000	-0.269
KCNMB1*	0.000	-0.259
ANGPT1*	0.000	-0.256
ATP1A2*	0.000	-0.244
SPARCL1*	0.000	-0.230
GAS1*	0.000	-0.228
TRIM29*	0.000	-0.212
ZIC2*	0.000	0.185
DPT*	0.000	-0.174
MEIS2*	0.000	-0.159
SLC22A3*	0.000	-0.156
SEMA3C*	0.000	-0.143
TPX2*	0.000	0.138
TOP2A*	0.000	0.129
CENPF	0.000	0.108
EGR1*	0.000	-0.100

3.4.3 Numerical Evaluation of the Dimension Reduction For Visualization

The second application of MetaPCA is direct extension of the first application, more correctly speaking it is a numerical evaluation of previous low dimensional reduction for visualization. The question of the first application is how to measure the informative gain in a numerical form. Although the cyclic pattern in cell cycle data is hard to evaluate quantitatively, we can measure the information gain or loss by using some machine learning techniques in some other cases. For this specific aim, we have used prostate cancer data and mouse metabolism data which both have multi-class subjects. We applied the same procedure as the first application—2D dimension reduction—, but instead of plotting results, we calculated median silhouette width of subjects assuming the three class labels are composed of real clusters. The rationale of this evaluation is based on the fact that an observer already anticipates for some patterns of subject distribution if the subject classes are already known as we can see in the cyclic pattern in the cell cycle data. Both prostate and metabolism data are composed of three known class labels, so we could conclude that the good dimension reduction for visualization should lead a good separation among those groups. Although silhouette width is utilized usually in the clustering evaluation purpose, we can use it to quantify the class separation in a robust way. The silhouette width measure can be defined as follows:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

, where $a(i)$ represents the average dissimilarity of subject i to all other objects of the assigned cluster (A), and $b(i)$ represents the minimum dissimilarity of the subject i to a subject in the other cluster (C) such that $C \neq A$. The range of silhouette width, $s(i)$, is between -1 and 1. Zero width means the membership of the subject is in the border line. Larger width means the better separation between classes.

The figure 11 shows the nice property of MetaPCA very well. Red dots represents the cases when the driven meta subspace incorporates evaluation dataset, and blue dots means the subspace did not utilize evaluated data. We can say that Heart and Skeleton data have better results in their own space, and the performance does not hurt much by combining with other data if their data themselves were used in the analysis, meaning red dots tend to

have higher scores. In the other hand, the results of Liver and Fat improved a lot by using PC subspace of others; especially if we compare two red dots in x = 1 or 4 in both Liver and Fat cases, we can see the results improved about 3 times.

The figure 12 shows the result of prostate cancer example. Compared to previous figure, overall MetaPCA performs did not improve much the outcome. The possible reason is those data set is more heterogeneous so that information gain from other subspace was smaller than the information loss, specifically Yu and Tomlins cases.

One of the common phenomena in both examples was the information gain experienced in the single subspace projection (x=1) was a possible barometer for the performance of MetaPCA. Specifically, the common result of Liver and Fat in metabolism data and Lapointe and Varambally in prostate data was that they had better results in the projection to other subspaces (See the dots in x=1, blue dots tend to be higher than the red dot). In this case, MetaPCA performed better. We can compare the two red dots in x=1 and 4; the red dot in x=4 was higher than the one in x=1.

This evaluation was a stepping stone to guide us to more fundamental question regarding to supervised machine learning. Although silhouette width could be served as a quantitative measure for class separation, we needed more sophisticated evaluation procedure to confirm that the common PC subspace could lead a better classification.

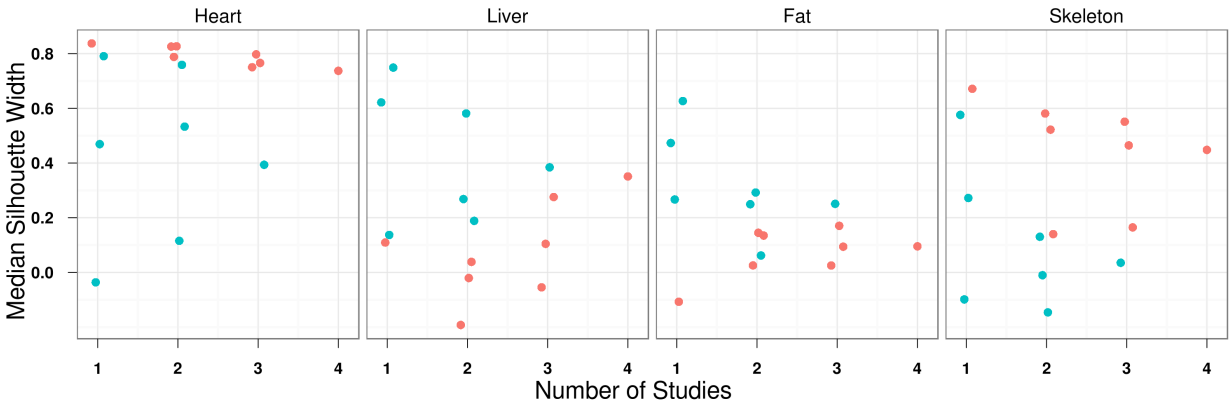


Figure 11: **Silhouette score for MetaPCA in mouse metabolism data.** Four mouse metabolism data were projected to the 2D meta-subspace found by eigenvalue maximization method. Red dots represents the cases when the driven meta subspace incorporates evaluation dataset, and blue dots means the subspace did not utilize evaluated data. X axis represents the number of studies included in each subspace generation. From 2 to 4 are based on MetaPCA, 1 is from single study. Y axis represents median silhouette width, which measures the cluster tightness or separation.

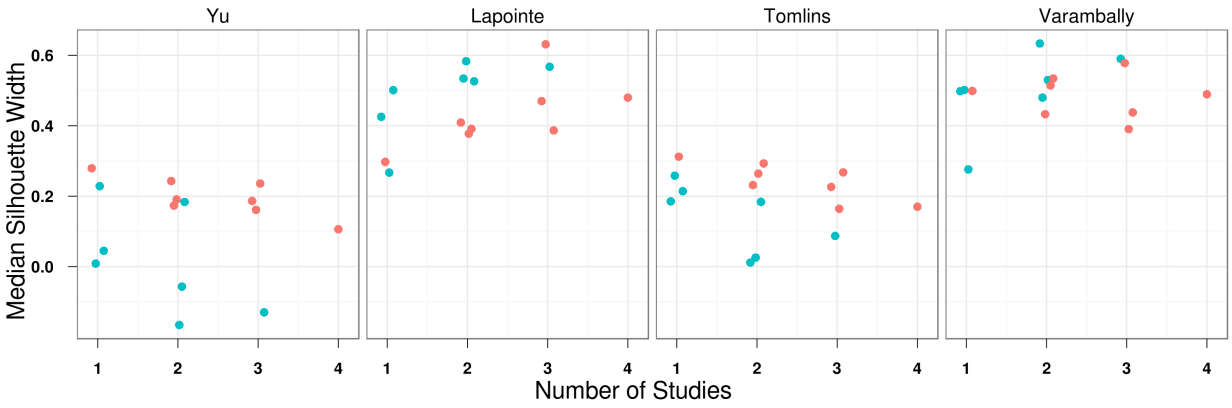


Figure 12: **Silhouette score for MetaPCA in prostate cancer data.** Four prostate studies were projected to the 2D meta-subspace found by Meta-SPCA. Red dots represents the cases when the driven meta subspace incorporates evaluation dataset, and blue dots means the subspace did not utilize evaluated data. X axis represents the number of studies included in each subspace generation. From 2 to 4 are based on MetaPCA, 1 is from single study. Y axis represents median silhouette width, which measures the cluster tightness or separation.

3.4.4 Supervised Machine Learning (Classification)

The third application of MetaPCA is to see if the new common subspace between data set could lead to the fundamental advantage to classify subjects better than the usual individual classification approach. For this specific goal, we used 5 brain cancer data which compare Anaplastic Astrocytoma (AA) and Glioblastoma multiforme (GBM) and 5 prostate cancer data which compare normal and primary subjects. To obtain more reliable and robust results, we have adopted leave-one-out cross-validation to test each test sample, specifically we developed classifier based on learning samples without a test sample in each iteration and meta-subspace was also re-generated based on learning samples. As the classification method, we have utilized shrunken centroids regularized discriminant analysis (SCRDA) [36], which produced reliable results with both original variables (genes) and derived PCs as features. For the feature selection, we depended on the gene selection procedure of SCRDA when we used genes as features; for the decision of number of PCs as features, we have used the first 5 PCs and we also let SCRDA to find the best PC for each learning procedure. For the evaluation of classification, we utilized Youden index [121], which can be calculated as $Sensitivity + Specificity - 1$.

The figure 13 shows the result from the five brain cancer data. Overall, we observed quite amount of improvement in accuracy for the classification, specifically Freije, Phillips, and Sun were among the best results we had. In the figure, x=2 to x=5 represent the results from MetaPCA; x=1 represents the direct projection to the individual data PC space; lastly x=0 was incorporated for the comparison purpose when we have used the original variables (genes) as features. Although the results of Gravendeel and Petalidis were not as good as the others, their results still show that MetaPCA accomplished similar performance as the single direct projection.

The figure 14 shows the result from the five prostate cancer data. Compared to brain data, most studies had already greater classification accuracy individually—one of the possible reasons is that the primary cancer subjects were easy to classify from normal subjects. As we saw in the figure 12, many studies were not well performed in the other subspace than their own subspace (See red dots are higher than blue dots); especially, in the cases of

Varambally, Wallace, and Yu, MetaPCA was performed somewhat less than the individual PCA in its own subspace (i.e. red dot in x=1). However, the interesting observation which we already observed in the previous section is that their direct individual projection to other studies (i.e. blue dots in x=1) was already not good. As we mentioned before that the results of the direct individual projection (i.e. when x=1) are determinants of the performance of MetaPCA. The implication of this result is that we need more homogeneous studies in a meta-analysis for better results.

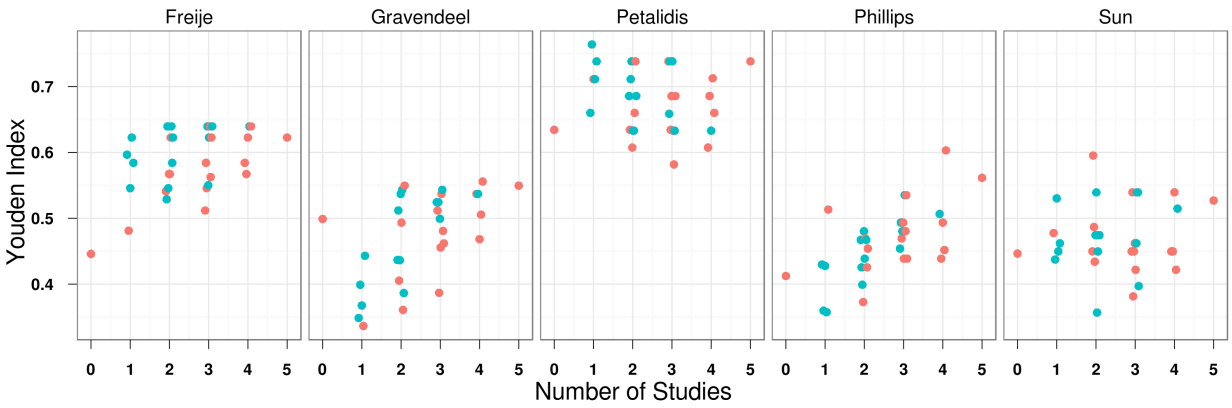


Figure 13: **Youden index for MetaPCA in brain cancer data.** Five brain cancer studies were projected to the meta-subspace found by Meta-SPCA with 100 genes for each PC. Red dots represents the cases when the driven meta subspace incorporates evaluation dataset, and blue dots means the subspace did not utilize evaluated data. X axis represents the number of studies included in each subspace generation. From 2 to 5 are based on MetaPCA, 1 is from single study. Zero represents the case when classification is based on original features, not on PCs. Y axis represents Youden index, which can be calculated as sensitivity + specificity - 1

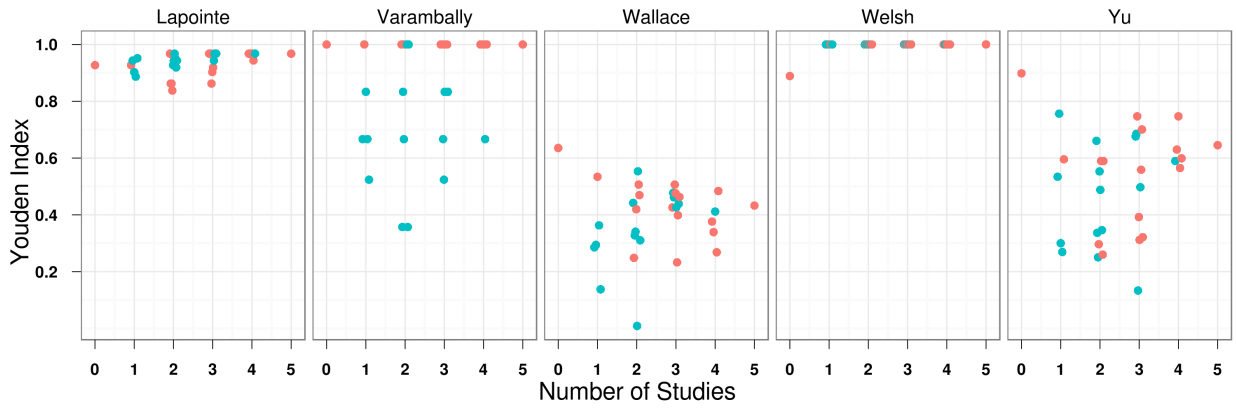


Figure 14: **Youden index for MetaPCA in prostate cancer data.** Five prostate cancer studies were projected to the meta-subspace found by eigenvalue maximization method. Red dots represents the cases when the driven meta subspace incorporates evaluation dataset, and blue dots means the subspace did not utilize evaluated data. X axis represents the number of studies included in each subspace generation. From 2 to 5 are based on MetaPCA, 1 is from single study. Zero represents the case when classification is based on original features, not on PCs. Y axis represents Youden index, which can be calculated as sensitivity + specificity - 1

3.5 SIMULATIONS

We have also executed a couple of simulation studies to investigate comparative advantages among our proposed methods and accuracy of important feature selection in Meta-SPCA. For random cluster generation, we used the approach proposed by Qui and Joe [82] which incoporates a cluster separation index measure. It also enables noisy variables and outlier samples to be imposed for mimicking the complexity of real data sets. At first, we simulated 100 full data sets which has 300 samples from 3 clusters and 2100 gene features out of which 2000 are noise, and additionally we generated outlier samples from the uniform distribution. We assumed 5 compatible studies, which were subsets of a full data set, had the same cluster size (20 samples in each cluster). In reality, each study is not directly comparable or combinable as in this case, but we simplified the simulation without considering further perturbation in each study level; we can consider the direct combination of 5 subsets as the best integration which leads to the most information gain. We summarized the mean of all 100 simulations of each category.

Figure 15 shows the simulation results including two basic MetaPCA methods ('Angle' and 'Eigen'). As a comparison, PCA results in an individual study ('Indiv') and the maximum possible case ('Max'), which includes all available samples and excludes noise features, are shown. We used median silhouette width as the performance measure which is represented as y-axis. X-axis represents the degree of cluster separation which is defined as $J(a) = \frac{L_2 - U_1}{U_2 - L_1}$, where L_k and U_k denote the lower and upper $\alpha/2$ percentile of projected cluster k, respectively. It is expected that the larger cluster separation index is, the larger median silhouette width is. Figure 15 shows the 9 results by different simulation settings in the amount of noise features and outlier samples. As expected, overall patterns are toward the upper right-hand side reflecting the very strong positive correlation between cluster separation index and silhouette width. The PCA result using all samples shows the best possible performance which is located above all others in every circumstances. Individual PCA shows its lack of power because of its smaller sample size; it locates in the bottom in most of the cases. The two basic MetaPCA methods performed similar in between both usual PCA cases in non-outlier cases (See the top row in Figure 15). The angle method performed

well consistently regardless of the existence of outliers and noisy features; instead, the eigen method suffered when there were outliers, particularly in the case when clusters are not well separated (See the middle and bottom row in Figure 15). We can conclude that the angle method is more robust than the eigen method to outlier samples and noisy features, and the angle method is expected to be more practical in most of real situations. The result was one of basis that we chose to extend only the angle method to incorporate Robust PCA and Sparse PCA. Additionally, results of the eigen method were almost the same as the ones of usual PCA approach with all sample as well as noise features (results are not shown).

Figure 16 shows the simulation results including two extended MetaPCA methods ('RobustAngle' and 'SparseAngle'). As a comparison, the angle minimization MetaPCA ('Angle') and maximum possible study ('Max') are shown. The robust angle method showed enough robustness to noisy features and outlier samples (See the middle and bottom row in Figure 16). However, interestingly, the angle method outperformed the robust method in terms of robustness as well as efficiency—the angle method had more stable performance in most of the regions and better in most of realistic situations. The fact that the angle method outshined the robust angle method in terms of robustness tells that the angle method has already the innate robustness to outliers so that in most of real applications additional robust procedure is not necessary. On the other hand, the results of Meta-SPCA method were showed to be sensitive to outliers. Considering its superb performance in non-outlier cases (See the top row in Figure 16), its dramatic degradation of performance in the existence of outliers is alarming; Sparse PCA methods based on the usual SVD technique should be used carefully and robust SVD procedure should be incorporated in the case.

Figure 17 shows the effectiveness of Meta-SPCA to find the true features. This simulation assumed no existence of outlier samples. We used adjusted rand index (ARI) as the accuracy evaluation measure of feature selection, which is defined as $ARI = \frac{RI_{obs} - E(RI)}{RI_{max} - E(RI)}$, where RI_{obs}, RI_{max}, and $E(RI)$ are observed, maximum, and expected rand index, respectively, and $RI = \frac{\text{sum of concordant pairs}}{\text{total number of pairs}}$. The best possible accuracy of feature selection was calculated by Sparse PCA using the full data set, which is represented as dashed blue line with squares located at the top in every cases. As another comparison purpose, the result of Sparse PCA using individual subset of the data was represented dotted green line with triangles which

are located at the bottom in most cases. The result of Meta-SPCA was represented as solid red lines with circles. Overall conclusion of Figure 16 is that with proper setting of the penalty parameter (λ), the true features can be chosen accurately without imposing many noise features.

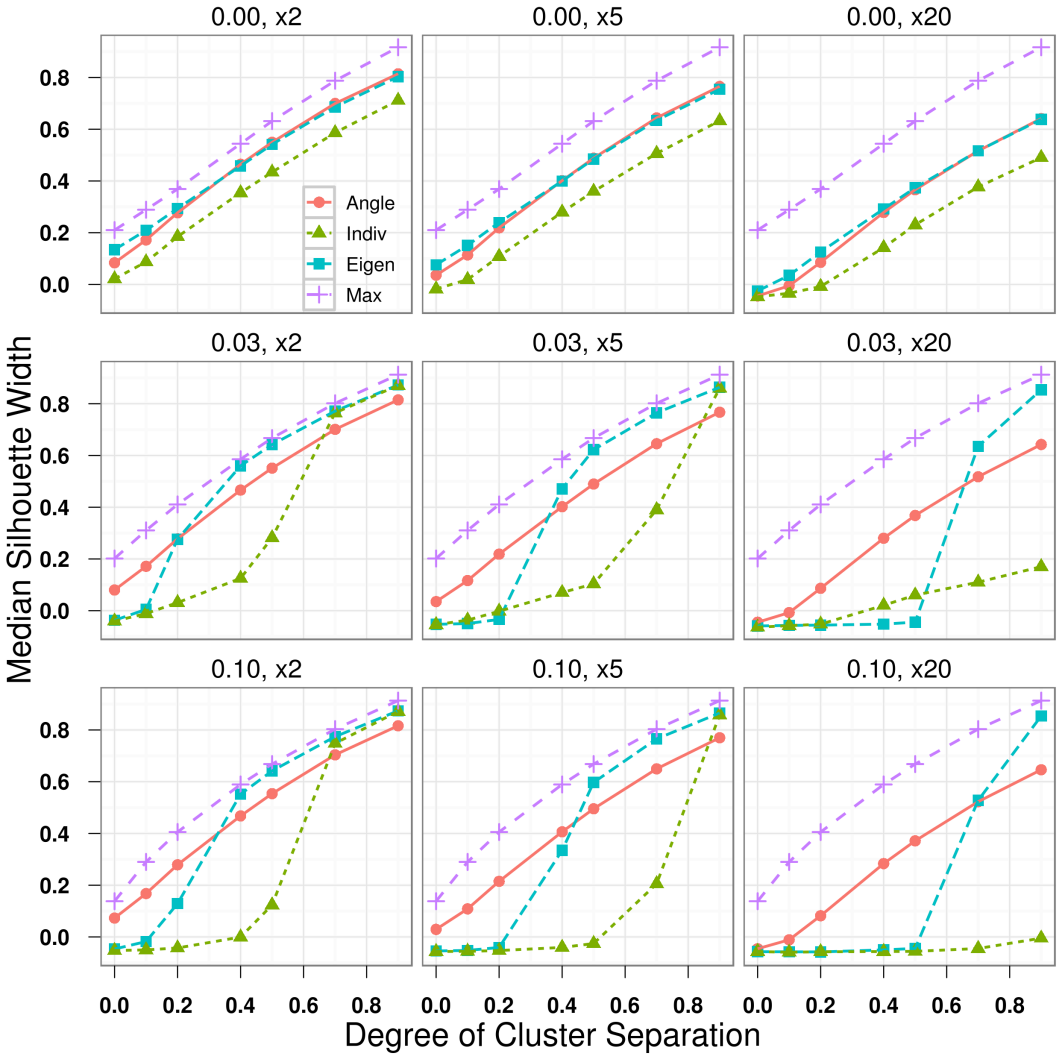


Figure 15: **Simulation results including the two basic MetaPCA methods.** Y-axis represents median silhouette width of true cluster samples. X-axis represents degree of cluster separation. Each plot represents the result of a case which specifies the proportion of outlier samples and noise features, e.g. the right bottom plot shows the result when there exist 10% outliers and 20 times of noise features. Each line represents the result of four analyses: maximum possible pca performance ('Max'), two basic MetaPCA methods ('Angle' and 'Eigen'), and individual pca performance ('Indiv'). Each point represents the mean of 100 simulations.

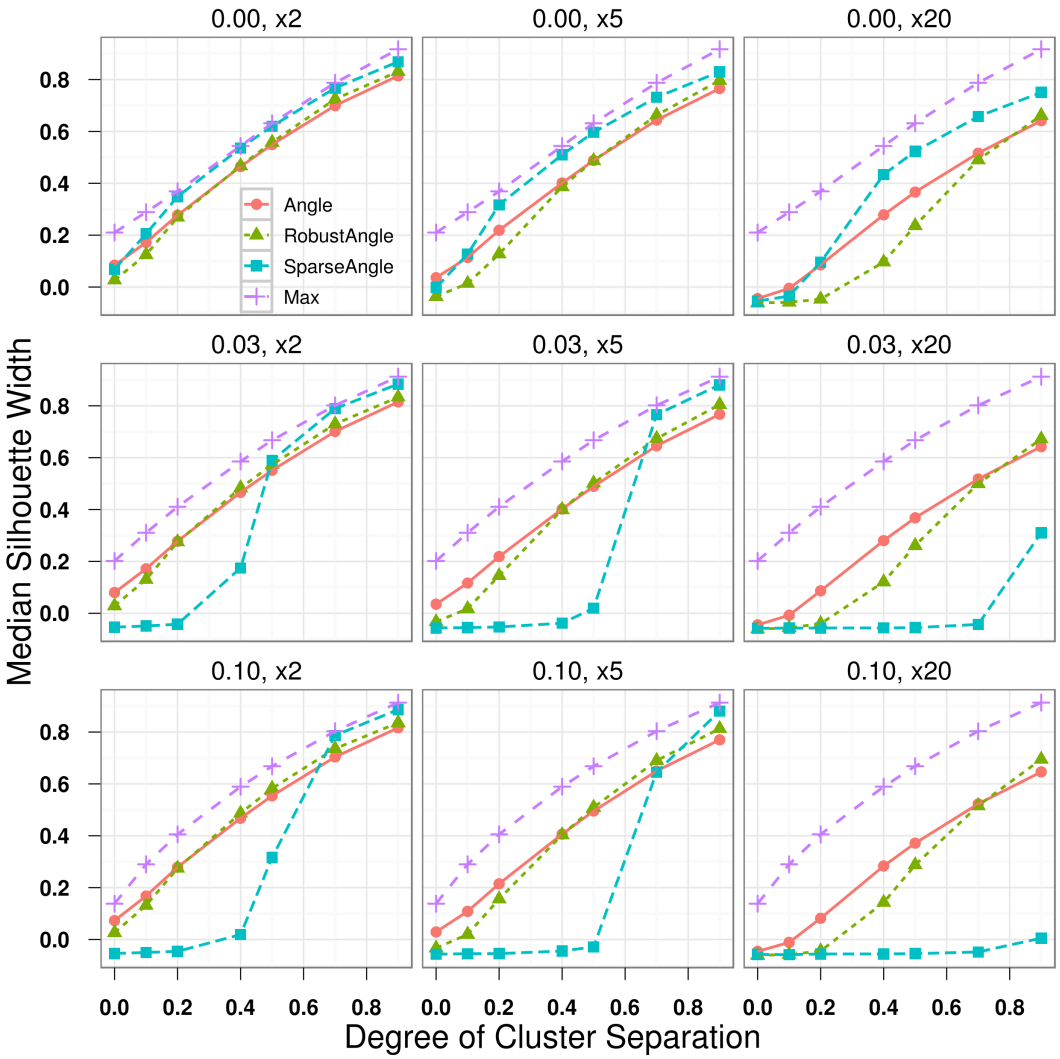


Figure 16: **Simulation results including the two extended MetaPCA methods.** Y-axis represents median silhouette width of true cluster samples. X-axis represents degree of cluster separation. Each plot represents the result of a case which specifies the proportion of outlier samples and noise features, e.g. the right bottom plot shows the result when there exist 10% outliers and 20 times of noise features. Each line represents the result of four analyses: maximum possible pca performance ('Max'), two extended MetaPCA methods ('RobustAngle' and 'SparseAngle'), and a basic MetaPCA method ('Angle'). Each point represents the mean of 100 simulations.

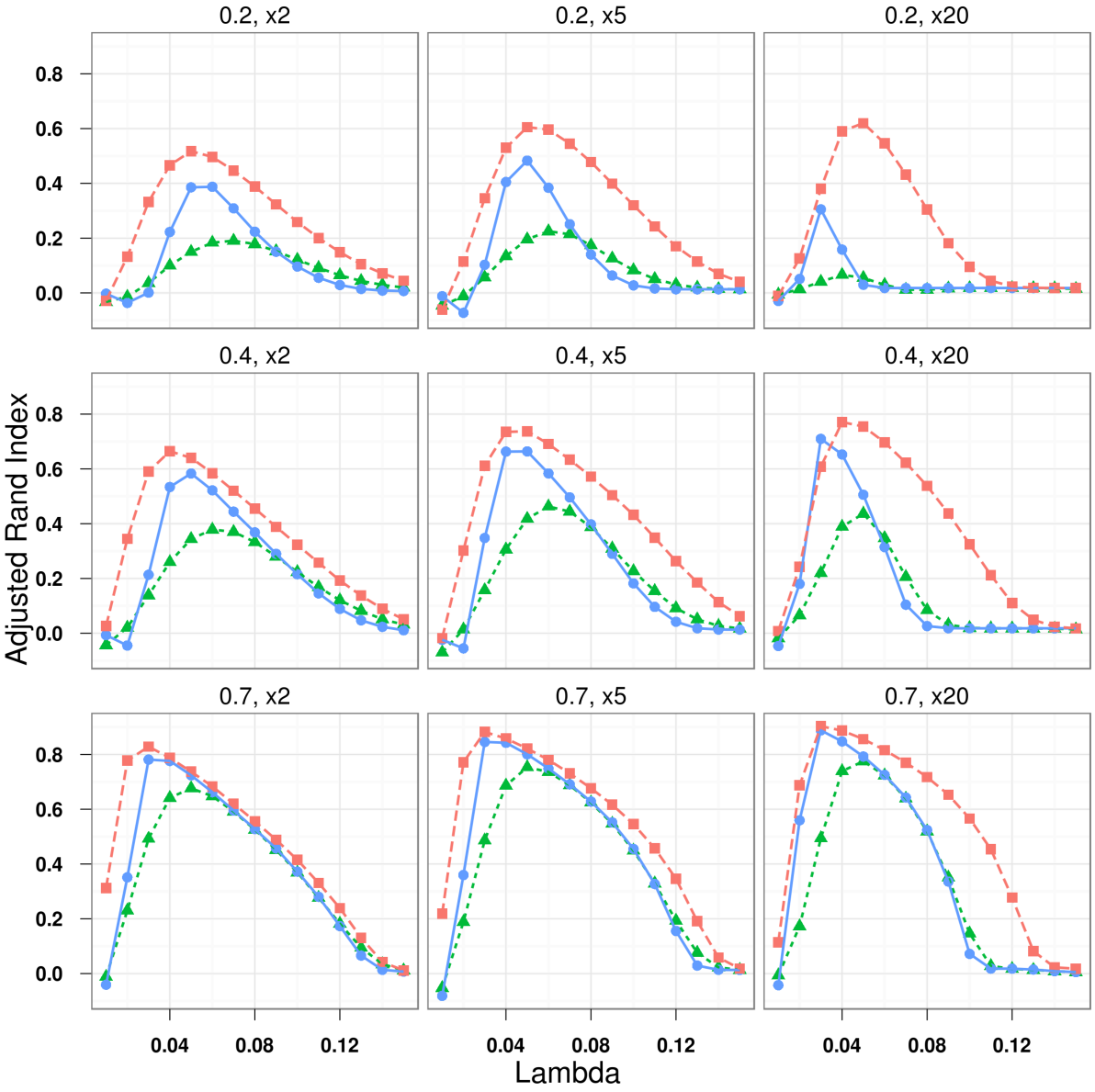


Figure 17: **Accuracy of feature selection in Sparse PCA and Meta-SPCA.** Y-axis represents adjusted rand index measuring the accuracy. X-axis represents different penalty parameter setting; the larger lambda is, the fewer features are selected. Each plot represents the result of a case which specifies the degree of cluster separation and the proportion of noise features, e.g. the right bottom plot shows the result when the clusters are very well separated (0.7), and there exist 20 times of noise features. Each line represents the result of three analyses: maximum possible SPCA performance ('Max'), Meta-SPCA, and the individual SPCA ('Indiv'). Each point represents the mean of 100 simulations.

3.6 DISCUSSIONS AND CONCLUSIONS

We consider the issue of information integration of multiple genomic (mainly microarray) studies for PCA analysis. When multiple microarray studies are available, each study can generate its own PC subspace. In this naïve approach, the information contained in multiple studies is not integrated and the projected data in each individual study are not comparable. An ideal integrative PCA approach is to integrate information in multiple studies and project data of all studies onto a single "optimal" PC subspace. In this paper, we have proposed and compared two MetaPCA approaches: eigenvalue maximization approach and angle minimization approach. We hypothesize that PCA results from multiple compatible studies used for integration have common cause of variation and underlying pattern regardless of individual noises and heterogeneity in the studies. Our methods focus on the commonness among studies to find the "optimal common projecting space". Similar concepts have been discussed in the statistical literature. For example, Krzanowski was the first to compare and combine PC subspaces across studies [56]. Neuenschwandera and Flury [70] discussed the theory underlying common PC models for dependent groups. These techniques, however, have not been extended and applied to genomic data analysis since existing microarray meta-analyses literature mostly focus on detection of differentially expressed genes or pathway enrichment analysis. To our knowledge, our paper is the first investigation of PCA application in microarray meta-analysis.

So far we have proposed several approaches to obtain common PC subspace and evaluated our methods in the context of dimension reduction for better visualization and classification. Although most proposed methods are based on existing ideas, our novelty is that we showed the information gain by combining multiple studies in several practical genomic data analyses. In our knowledge, there was no similar attempt for the same goal as ours in genomic research field yet.

From all the examples, we can observe the consistent patterns of result that make us to get a glimpse of the performance of MetaPCA. When a study can be represented better or quite similarly in the other studies' PC subspace as original subspace, we observed that the study could obtain more information gain so that it performed better in terms of visualization

and classification. The striking example of cdc15 in Spellman cell cycle data and Liver or Fat data in mouse metabolism are the best example of the argument. However, as we can see in the prostate example, MetaPCA does not guarantee better performance all the time, sometimes it cost more than benefits. Wallace or Yu in the classification are the examples of such cases. Prostate cancers were reportedly very heterogeneous disease than other cancers [89]; prostate cancer data may not be a good data for a meta-analysis. Although it is still controversial that if all studies should be included in a meta-analysis or only the homogeneous ones should be [23], our results support homogeneous study selection is very important pre-process for the success of meta-analysis. In this context, our novel MetaQC [50] for genomic meta-analysis could be a good tool to screen out inappropriate studies before meta-analysis procedure.

4.0 UNSUPERVISED COMBINATION OF CLINICAL AND GENE EXPRESSION DATA ELUCIDATES PHENOTYPES IN ILD AND COPD

(It is expected to be published in a high profile medical journal)

It is necessary to define phenotypes of disease states to investigate their underlying molecular mechanisms and develop new treatment strategies, therapeutics, and biomarkers. Using chronic lung diseases which are commonly thought of as clinically distinct, we demonstrate a computational approach to reverse phenotype patients using both clinical and gene expression data. We acquired lung tissue, computed tomography, and clinical data on 472 subjects who were initially given a clinical diagnosis of either interstitial lung disease (ILD) or chronic obstructive pulmonary disease (COPD). We performed unsupervised clustering on patients with both phenomic and genomic expression data. We developed insightful feature visualization tools to explore and interpret the clusters. Pathway analysis and clinical feature correlation are performed to characterize and annotate the identified intermediate phenotypic patients. We showed the convergence/divergence patterns of disease phenotypes in the integrated clustering by clinical and molecular features. Large number of patients was in off-diagonal clusters which represent discrepancy between clinical and molecular phenotypes. This is the first attempt that systematically integrate transcriptomic and phenomic data in ILD and COPD. We identified new clusters of intermediate phenotypic patients that may lead to improved understanding of these diseases and novel therapeutic approaches. Our findings reflect that current clinical definitions and classification do not account for the large number of patients having intermediate phenotypes or less common features that are often excluded from clinical trials and epidemiology reports.

4.1 INTRODUCTION

Chronic lung diseases affect a significant portion of the population. Approximately 24 million adults in U.S. have evidence of abnormal lung function. They go through 9.5 million office and emergency room visits, 726,000 hospitalizations, and 119,000 deaths each year [66]. While the majority of these deaths can be attributed to chronic obstructive pulmonary disease (COPD), the major smoking induced lung disease, more than 15,000 can be attributed to idiopathic pulmonary fibrosis (IPF), a relentless, nearly always fatal fibrotic lung disease also associated with smoking [73, 1].

The phenotypic description of these chronic lung disorders has been the focus of interest of clinical research in the last decades of the 20th century. Clinical researchers dedicated significant efforts to define and identify the purest phenotypes of chronic lung diseases based on physiology, radiology features, histopathology, significant negative findings and most importantly prognosis [98, 32]. Such disease classifications create a unified vocabulary of lung diseases that allows better communication between clinicians and researchers and simplifies recruitment to clinical studies. While critically important and widely accepted, these clinical definitions and classifications do not account for the large number of patients that present with intermediate phenotypes or less common features. Such patients are often excluded from clinical trials and epidemiology reports, a justified practice that limits our ability to document the complexity and potential overlap of pathogenic processes and disease manifestations shared by emphysema/COPD and IPF. Importantly, existing classifications do not reflect recent important advances in radiologic imaging analysis and high-throughput gene expression methods, techniques that potentiate investigators to fully capture the complexity of a given individual's phenotype.

High-throughput gene expression methods have been widely applied in cancer research to identify known disease phenotypes or to define new ones. In pulmonary medicine, these approaches have not been as widely applied, however recent studies evaluating IPF, COPD or other lung diseases provide significant insight and promise in the application of similar techniques to large cohorts with lung disease transcending usual disease boundaries [75, 81, 90, 91, 100, 116, 117, 119, 124]. While none of these studies directly compare IPF

and emphysema/COPD and their subclasses, they do provide evidence that, in contrast to the prevailing paradigm that places dysregulated activation of matrix degrading proteases underlying emphysema in one extreme, and relentless extracellular matrix deposition in IPF at the other extreme, these diseases may share activation of similar pathways, and potentially similar mechanisms. While the investigators on this application completely agree that, in their extremes, emphysema and IPF likely represent different and divergent anatomical and temporal responses to injury, we hypothesize that by applying gene expression profiling and advanced computational approaches to a large enough and well characterized cohort of samples of IPF and emphysema/COPD, we will identify disease-relevant gene expression modules that are highly distinct, reproducible and characteristic of disease phenotypes that go beyond current disease definitions.

To address our underlying hypothesis that gene expression patterns will identify diverging and converging phenotypes in known phenotypes in IPF and COPD, we show gene expression signatures that globally characterize COPD and IPF. Modular integration of data from multiple sources using gene expression, clinical data, radiological and physiological data may reduce the effects of individual variation. The significance of our works is the innovation that looks at IPF and COPD, not as phenotypic extremes, but as multiple syndromes that may be the end result of overlapping as well as diverging mechanisms. The availability of the molecular phenotypes within and across disease boundaries will have the potential to liberate pulmonary clinical research from the need to focus on the most divergent phenotype, and instead will allow researchers to focus on mechanistically relevant disease molecular phenotypes.

4.2 MATERIALS AND METHODS

We acquired flash frozen lung tissue, computed tomography (CT) data, and clinical data on 472 subjects from the Lung Tissue Research Consortium (LTRC). These subjects were initially given a clinical diagnosis of either interstitial lung disease (ILD) or chronic obstructive pulmonary disease (COPD) based on their clinical, pathologic, and radiographic data (217

Table 12: Correlation of different data types

	Quantitative	Ordinal	Nominal
Quantitative	Pearson	Biserial/Polyserial	Point Biserial/MCC[1]
Ordinal	Biserial/Polyserial	Spearman/Tetrachoric	Rank Biserial
Nominal	Point Biserial/MCC[1]	Rank Biserial	Phi/PCC[2]

[1] Multivariate correlation coefficient
[2] Pearson's contingency coefficient

COPD and 255 ILD). Extensive clinical variables (∼1,000) were obtained by questionnaires (demographic, medical history, family history, smoking history, concomitant therapy, symptom, SF-12 health, St. Georges respiratory, environmental, occupational), tests (six-minute walk test, cardiopulmonary exercise test, PFT, blood test, CT scan), and diagnosis reports (central and local pathology, clinical report). Gene expression data were obtained through the use of the two different Agilent microarray platforms. We applied loess normalization after matching probe ids between platforms. The number of matched genes are 15,966.

We filtered out clinical variables to choose informative and representative ones in each category (i.e. less missing values, closer to continuous). We defined the distance between variables as $1 - \rho_{ij}$, where ρ_{ij} is the correlation coefficient between two variables i and j. To calculate the correlation between different data types, we applied several correlation models as in the table 12. After several iterations, we selected 30 clinical variables.

We applied k-means sparse clustering [114] with 3 clusters to find informative genes. We obtained 4,291 genes which have positive contribution. We applied k-medoids [106] with k=3 using either clinical variables or genes as features. For an effective representation of the convergence/divergence patterns of clinical and molecular phenotypes, we have developed 2D or 1D visualization tools which show the transition of features of a cluster in the intuitive and interpretable way. We have modified GEDI plot [20] to incorporate clinical features and gene expressions in the same space using multi-dimensional scaling (MDS) [5, 88]. To find the common and augmented feature space, we combined clinical variables and genes

so that we obtained a 4,321×4,321 dissimilarity matrix and applied MDS to find the best 2 or 1 dimensional configuration of features. After 2D projection, we divided the derived space to 10x10 grid subspace and calculated median standardized score of the cell (clinical or genes) for each patient. For cluster visualization, we summarized medians of each grid of given cluster members. For 1D projection, we applied spaghetti plots that fit a cubic spline to standardized scores of each cluster subjects (clinical or genes). Since the similar features are closely located, we can interpret each figure based on a cluster of features. Moreover clinical variables and genes can be used together for the interpretation interactively. Principal component analysis (PCA) was applied to project subjects in 2 dimensional clinical feature space. To enhance interpretation, we juxtaposed clinical features on top of derived PC subspace (PCA biplot). The coordinate of a feature corresponds to the amount of correlation that the feature has with each derived principal component.

For validation of cluster analysis, we divided samples to two groups. The training group was composed of the first two batches (305 samples), and the testing group was the third batch (167 samples). Feature filtering was applied to the training set using k-means sparse clustering [114] with k=3, and 4402 genes were selected (30 clinical variables were reused). Aforementioned cluster analysis and visualization methods were applied to the training set. We applied a separate cluster analysis to the testing set using the same features found by training set. To validate the testing set result, we utilized the 2D and 1D cluster visualization of training set as the gold standard, i.e. we reused the feature projection result of training set to visualize the clusters found by the testing set. The rationale of this validation study is that if the clustering result found by training set was not a coincidence, it should be reproducible with independent data.

4.3 RESULTS

4.3.1 Clinical Feature Filtering

We applied clinical feature visualization and variable clustering method to find representative variables using the distance matrix calculated by various correlation models in Table 12. Figure 18 shows an example of clinical feature visualization. Each cell represents the degree of pair-wise similarity between two variables, which is defined as $|cor_{ij}|$, where i and j represent two variables. We have generated many figures like Figure 18, which were served to help us to understand clinical data set we have and to promote us to communicate between physicians and statisticians. Using the same distance matrix, which is $1 - |cor_{ij}|$, we also applied a hierarchical clustering method to find clusters of variables such as Figure 19. We manually reviewed variables cluster by cluster and chose 1 or 2 clinically important representative variables in a cluster. We iteratively applied this approach and decided to include 30 informative and clinically important variables.

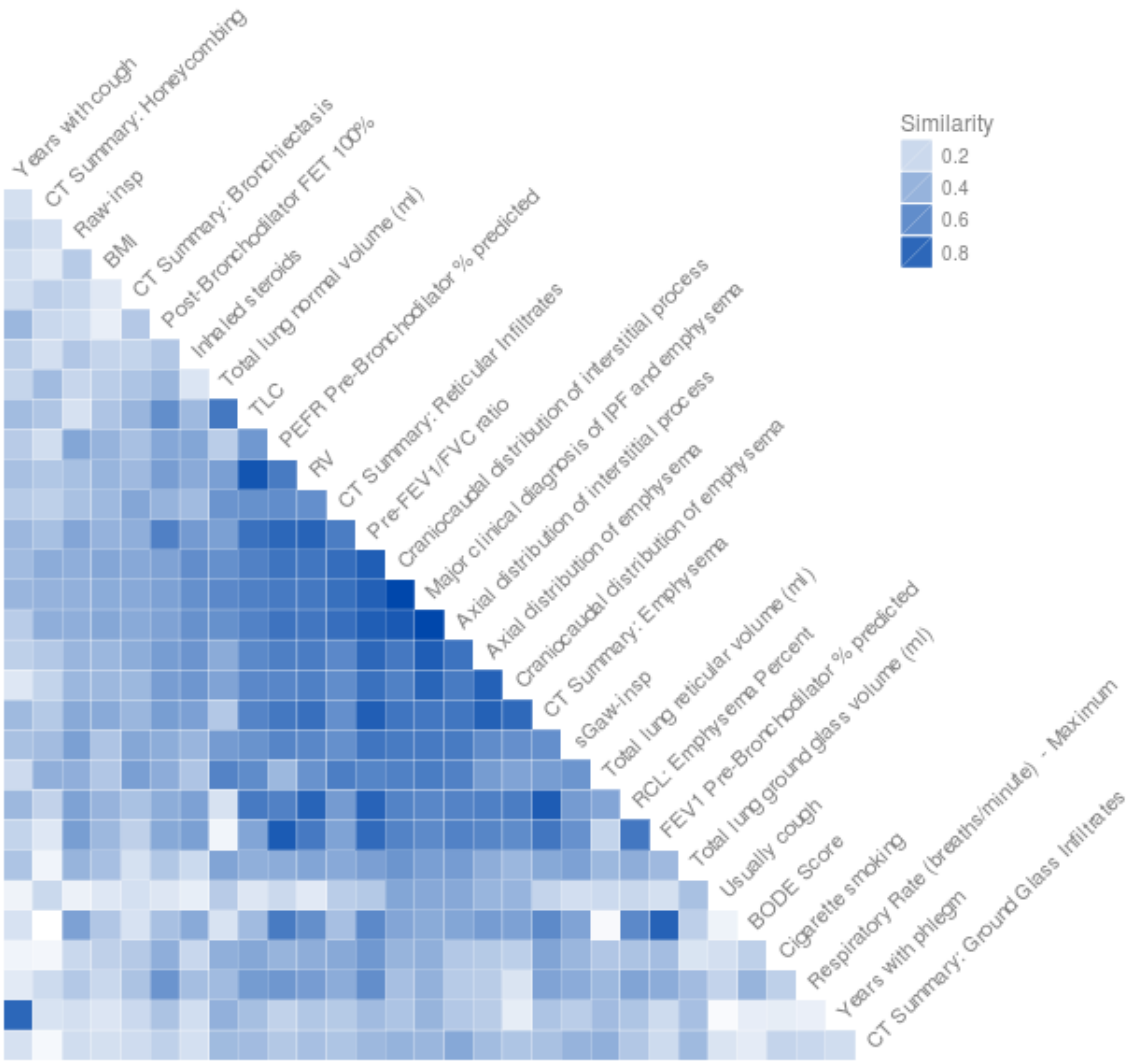


Figure 18: **An example of clinical feature visualization.** Each cell represents the degree of pair-wise similarity (correlation coefficient) between two variables.

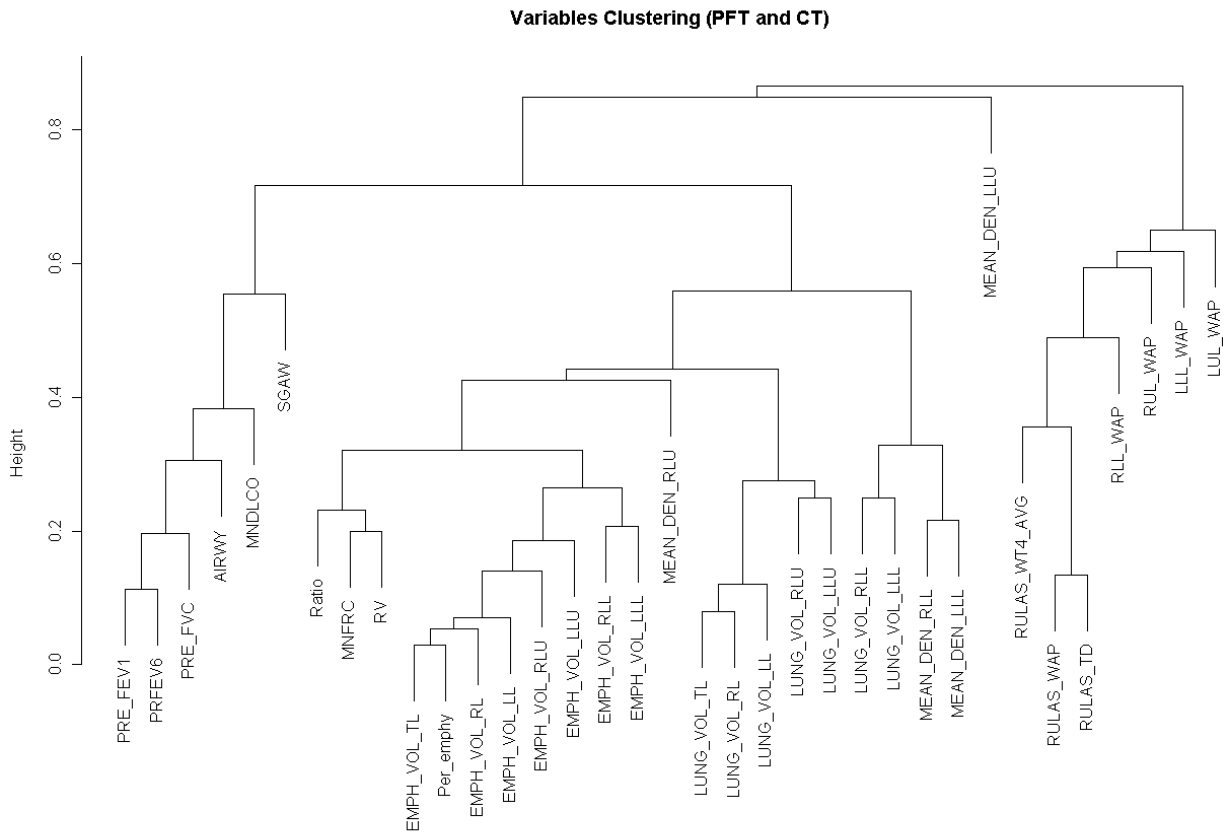


Figure 19: **An example of clinical variable clustering.**

4.3.2 Cluster Validation Analysis

One of concerns in pure computational approach is over-fitting. Although we didn't assume any models to fit, we still wanted to confirm that our methods are robust enough to reproduce the clusters we found. Here, the reproduction of clusters means the characteristics of each cluster can be reproduced in terms of clinical phenotypes and gene expression patterns. So we utilized our novel visualization tools to show the validation of clustering. Figure 20 shows the result of cluster validation analysis by 2 dimensional gene expression visualization. Each cell represents a cluster of patients which found by two separate cluster analyses using gene expression (x-axis) and clinical phenotypes (y-axis). Each pixel represents the median intensity of standardized gene expression in a group of genes similarly expressed. Red color represents over expression, and blue represents under expression. We observed testing set can reproduce the training set result very well as we can see the similar pattern of heatmaps between training and testing sets; specifically the clusters of genes in the left area are over-expressed in COPD clusters by gene expression, and the ones in the right area are over-expressed in ILD clusters. There was a little discrepancy in the middle column that in the training set the mixed cluster was closer to ILD cluster and in the testing set to COPD cluster; one of probable reasons can be found from the major clinical diagnosis composition of clusters in Figure 21 that the mixed cluster has more ILD proportion in the learning set and more COPD proportion in the testing set. Figure 22 represents the same cluster validation result using 1 dimensional gene expression visualization. In this figure, similarly expressed genes are located closely in the x-axis, and y-axis represents standardized gene expression intensity. Each line represents cubic spline estimate of gene expression in a cluster of patients which found by two separate cluster analyses using gene expression (three levels of vertical cells) and clinical phenotypes (three levels of different colors in a cell). As the 2D visualization, we could confirm that the cluster analysis based on training set is reproducible.

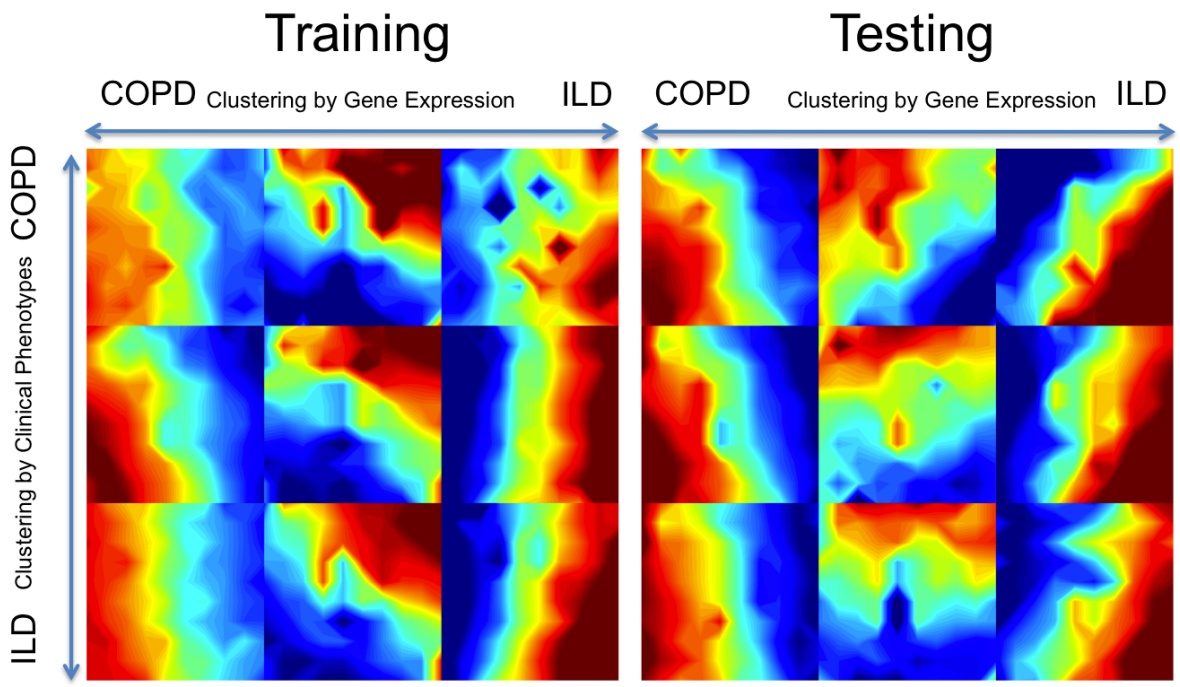


Figure 20: **Cluster validation by 2D gene expression visualization.** Each cell represents a cluster of patients which found by two separate cluster analyses using gene expression (x-axis) and clinical phenotypes (y-axis). Each pixel represents the median intensity of standardized gene expression in a group of genes similarly expressed. Red color represents over expression.

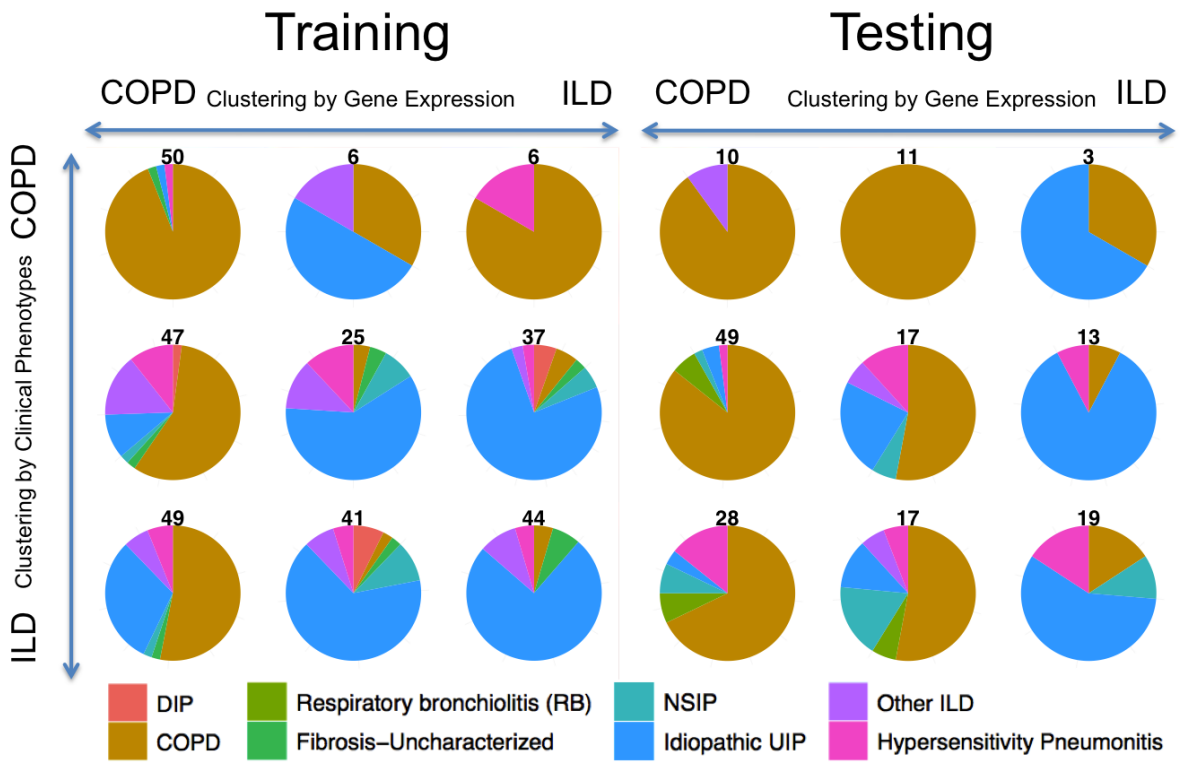


Figure 21: **Major clinical diagnosis in each cluster for cluster validation.** Each cell represents a cluster of patients which found by two separate cluster analyses using gene expression (x-axis) and clinical phenotypes (y-axis). Each pie chart represents the proportion of major clinical diagnosis in the patients of the specified cluster.

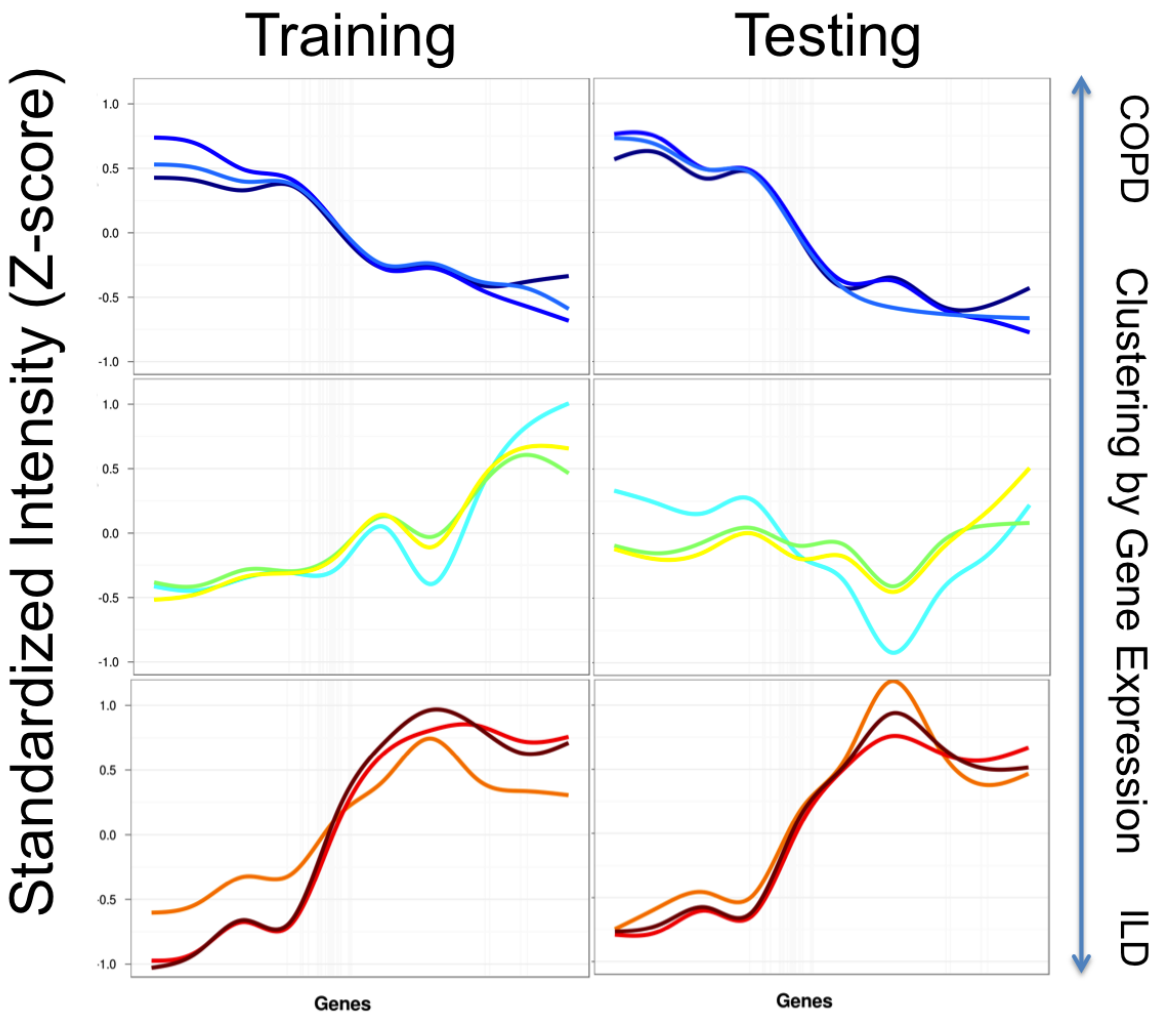


Figure 22: **Cluster validation by 1D gene expression visualization.** Each line represents cubic spline estimate of gene expression in a cluster of patients which found by two separate cluster analyses using gene expression (three levels of vertical cells) and clinical phenotypes (three levels of different colors in a cell). Y-axis represents standardized gene expression (z-score), and x-axis represents genes which are ordered by their pair-wise distance.

4.3.3 The Convergence/Divergence of Clinical and Molecular Phenotypes

One of main aims of this study is to show a transitional pattern of both clinical and molecular phenotypes so that we could focus on clusters which have discrepancy between those phenotypes. Figure 23 shows the change of phenotypic pattern among clusters: the plot a) represents pattern of gene expression that three clusters found by gene expression are differentiated clearly; however, the plot c), which represents change of clinical phenotypes, does not show stable and clear distinction among clusters. One of main reasons is the existence of many missing values in some patients and the sparsity of clinical variables in derived 2D space. Traditionally, only few clinical variables such as Pre FEV1 and the ratio of Pre FEV1/FVC were mainly used for those disease classification; however, we showed the necessity of incorporating gene expression as features for better classification. For instance, in b) cluster 1 and 2 are composed of similar proportion of ILD and COPD patients, and expectedly, in c) their clinical phenotypes are very similar; however, in a) gene expression pattern of two clusters are very different, suggesting that they are different diseases in spite of their similarity in clinical phenotypes. In this context, we need further study to find a correlation between their gene expression and disease progression or treatment efficacy.

Figure 24 shows one dimensional representation of the same information in Figure 23. The goal of this figure is to have integrated interpretation of clinical phenotypes and gene expression in a cluster. Both plots in the figure share the same x-axis, but y-axis applies only relevant scores which are standardized score in clinical variables in the above, and standardized expression intensity in the below. One example of integrated interpretation is that the clusters 3, 6, and 9 have similar gene expression patterns, especially they have important over-expressed set of genes—MMP11, COMP, MMP3, MMP14, MMP7, and MMP1—which are known IPF signature genes; moreover, those genes are closely related with important clinical phenotypes—Reticular infiltrates, Lung reticular volume, and ratio of Pre FEV1/FVC—which are known important CT and lung test features of IPF.

To our knowledge, it is the first time that a systematic sub-classification of two major chronic lung disease is done by using both clinical and molecular features with large number of well defined cohort.

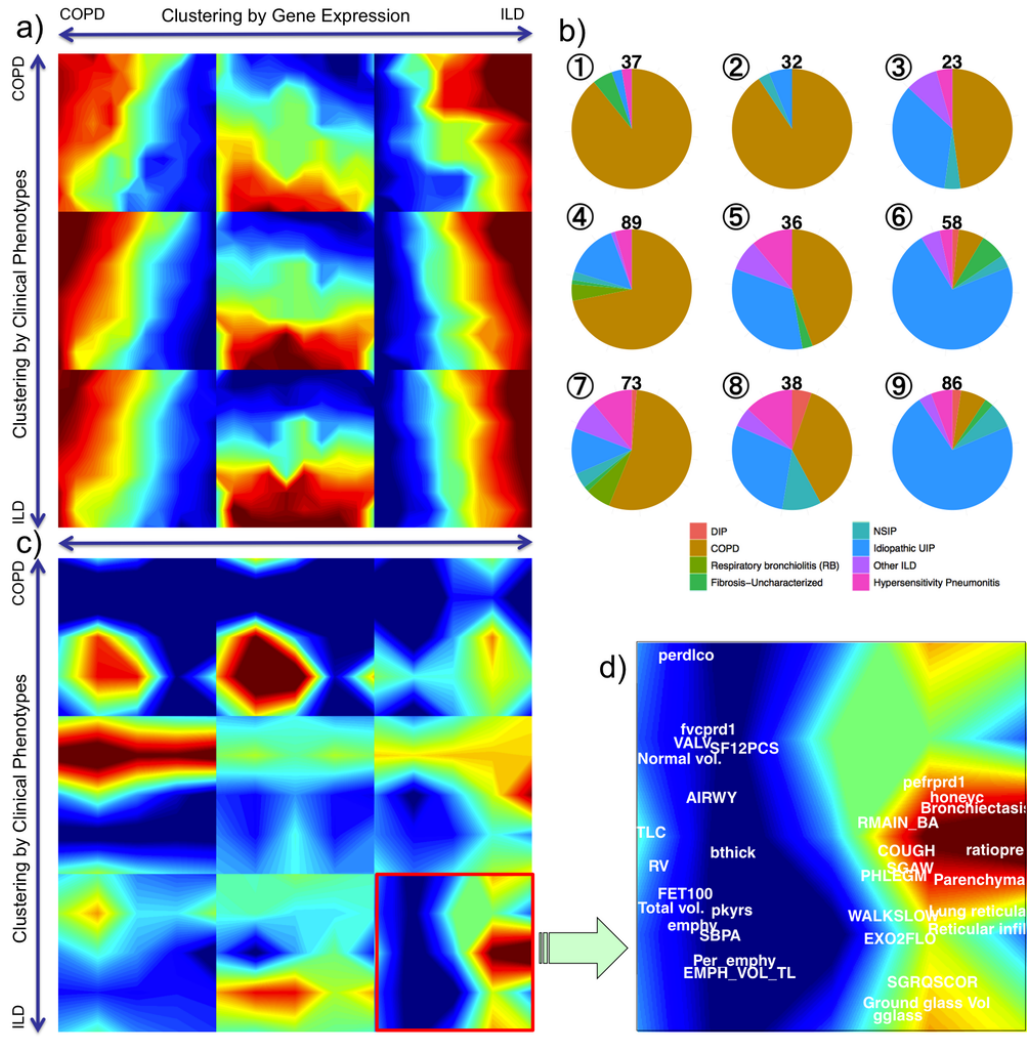


Figure 23: **The Convergence/Divergence of Clinical and Molecular Phenotypes.** a) Each cell represents a cluster of patients which found by two separate cluster analyses using gene expression (x-axis) and clinical phenotypes (y-axis). Each pixel represents the median intensity of standardized gene expression in a group of genes similarly expressed. Red color represents over expression. b) Each pie chart represents the proportion of major clinical diagnosis in the patients of the specified cluster. c) Similar to a), but each pixel represents the median value of standardized clinical variables nearby. d) The cluster 9 in c) was enlarged to show that the color patterns are interpretable by given clinical features juxtaposed on top of the figure in d). Since a) and c) are comparable, gene clusters in a) which are related to interesting clinical phenotypes in c) can be matched and interpretable interactively. (e.g. gene expressions and clinical phenotypes are highly correlated in the cluster 9, which has the largest IPF proportion)

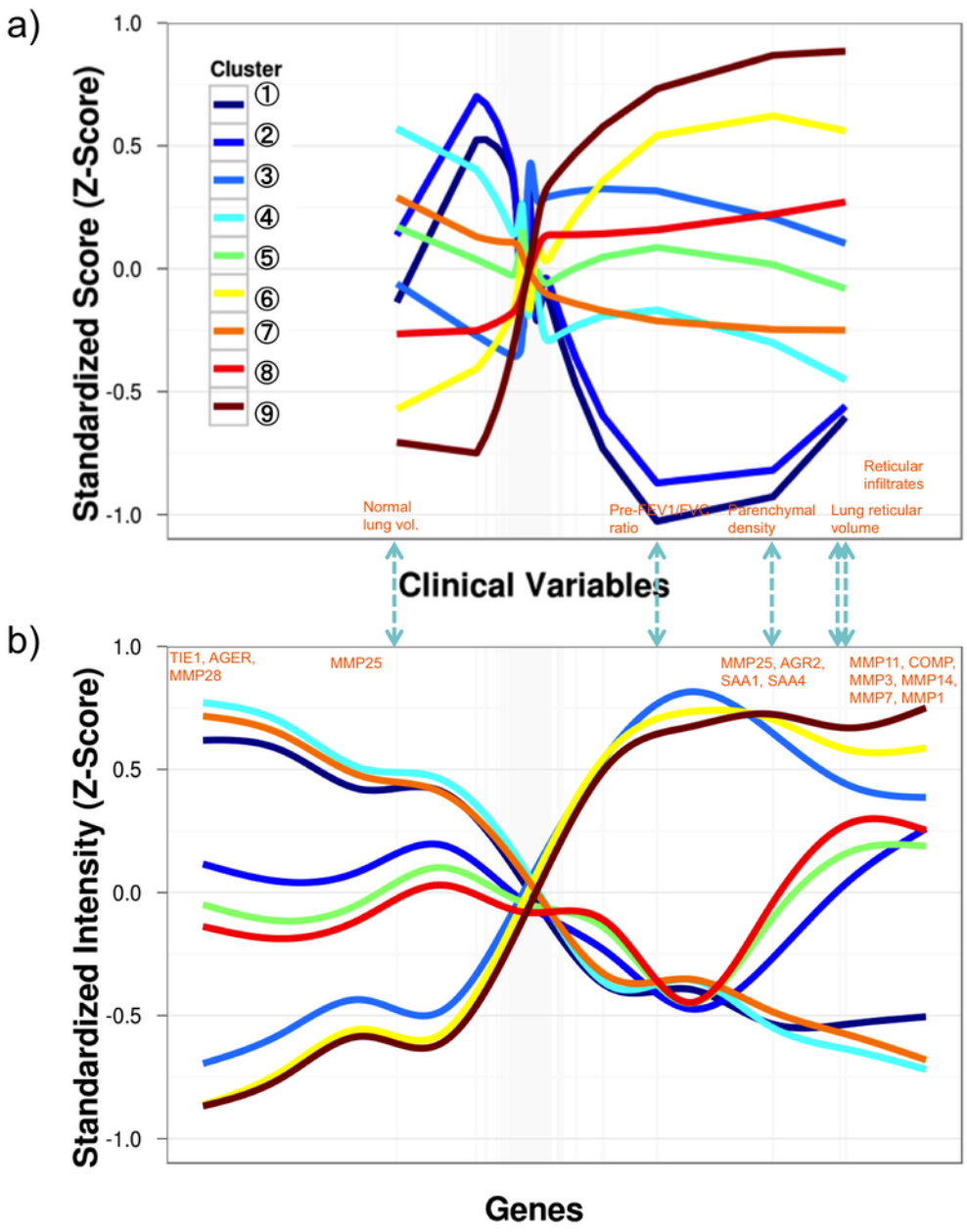


Figure 24: **One Dimensional Representation of the Figure 23.** Both x-axis's are exactly the same, derived by 1D projection of pair-wise distance structure using both clinical and molecular features. They are ordered by their similarity. a) In the x-axis, only clinical variables are presented for separate interpretation. Y-axis represents standardized value of each clinical variable. Each line represents a cubic spline estimate of scores in the specified cluster. b) Only genes are presented in the x-axis. Y-axis is the standardized gene expression. X-axis in both figures are comparable, so one can find a set of highly correlated reciprocal features.

4.4 DISCUSSION AND CONCLUSIONS

We showed in a systematic way that there exist groups of patients who have intermediate phenotypes and mixed molecular signatures between COPD and ILD, which are known as phenotypic extremes. It can be described as a phenomenon that both clinical and molecular phenotypes have different convergence/divergence patterns in each cluster. For instance, in Figure 23, clusters 1 and 9 have convergent pattern of clinical and molecular phenotypes fitting to current knowledge and disease definition: Clinical phenotypes were highly correlated to molecular phenotypes in "pure" disease groups. However, the existence of other clusters which have divergent pattern of clinical and molecular phenotypes support the argument that current disease definition is too narrow to incorporate intermediate phenotypic groups of patients. Particularly, cluster 1 and 2 have the similar diagnosis proportion and cluster sizes and turned out to have very different molecular signatures; cluster 7 has the most divergent pattern which has strikingly similar gene expression with the typical COPD cluster (cluster 1) in spite of its diverse composition of diagnoses. It implies that similar molecular processes can cause various clinical phenotypes that supports our hypothesis that ILD and COPD are not phenotypic extremes but multiple syndromes that may be the end result of overlapping as well as diverging mechanisms. For the first time, we showed that molecular phenotyping can be successfully incorporated with clinical phenotyping to identify a cluster of various homogeneous patients.

5.0 FINAL DISCUSSION AND FUTURE DIRECTION

We have investigated novel methodologies to show the advantages of proper genomic meta-analysis. Although study selection is necessary and justifiable practice, no known systematic and quantitative criteria exist. Our contribution is to reveal the pitfalls of inappropriate selection criteria and to offer a tool which can be used as an objective evidence of study inclusion/exclusion criteria. MetaPCA is a direct beneficiary of MetaQC. The success of simultaneous dimension reduction and discovery of effective common PC subspace depend on the homogeneity of given studies. Although our proposed approach is incorporated with robust methodologies, we observed homogeneous studies lead to better results. MetaPCA is a significant attempt to seek a new direction in genomic meta-analysis. We expect increased popularity of meta-analysis in various other machine learning techniques including meta clustering and meta discriminant analysis. The direct extension of our works is to apply them to different data types such as RNAseq and methylation data. Particularly, the data type in RNAseq is discrete instead of continuous which is the case in microarray data. More investigation is needed to check whether any assumption in our methods is affected. However, the philosophy and the goal in our proposed approaches still prevail.

The third component of this thesis was not about horizontal data integration but about vertical integration in that it was dealing with different data types exist in different levels. Over the last several decades, dominant disease definitions in chronic lung diseases were based on a few clinical phenotypes such as pre-FEV1 or the ratio of pre-FEV1 and pre-FVC. However, it has been well known that clinicians often faced atypical phenotypic manifestation in some patients which usually have both characteristics of ILD and COPD. There were no systematic efforts to reveal the phenotypic convergence/divergence using well defined large cohort. With the help of integrative clustering analysis and cluster visualization methods, we

showed and validated the existence of transitional pattern of disease phenotypes in clinical and gene expression data. The direct extension of our findings is to incorporate follow-up studies on the identified groups of patients in terms of disease progression and drug efficacy. After that, we could define novel subtypes of the two diseases using important clinical and molecular features which may lead to a novel prediction model or diagnostic tool.

BIBLIOGRAPHY

[1] K. Baumgartner, J. Samet, C. Stidley, T. Colby, and J. Waldron.
Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis.
American journal of respiratory and critical care medicine, 155(1):242, 1997.

[2] R. Bellman.
Adaptive control processes: a guided tour.
Princeton University Press, 1:2, 1961.

[3] Y. Benjamini and Y. Hochberg.
Controlling false discovery rate: a practical and powerful approach to multiple testing.
Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300, 1995.

[4] C. Bishop.
Pattern recognition and machine learning, volume 4.
Springer New York, 2006.

[5] I. Borg and P. Groenen.
Modern multidimensional scaling: Theory and applications.
Springer Verlag, 1997.

[6] R. Breitling and P. Herzyk.
Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data.
Journal of bioinformatics and computational biology, 3(5):1171–1190, 2005.

[7] A. Campain and Y. Yang.
Comparison study of microarray meta-analysis methods.
BMC bioinformatics, 11(1):408, 2010.

[8] E. Candes, X. Li, Y. Ma, and J. Wright.
Robust principal component analysis?
Arxiv preprint arXiv:0912.3599, 2009.

[9] J. Choi, U. Yu, S. Kim, and O. Yoo.
Combining multiple microarray studies and modeling interstudy variation.
Bioinformatics, 19(Suppl 1):i84, 2003.

[10] D. Cox and D. Hinkley.
Theoretical Statistics. 1974.

[11] M. Crescenzi and A. Giuliani.
The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data.
Febs Letters, 507(1):114–118, 2001.

[12] C. Croux, P. Filzmoser, and M. Oliveira.
Algorithms for projection-pursuit robust principal component analysis.
Chemometrics and Intelligent Laboratory Systems, 87(2):218–225, 2007.

[13] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet.
A direct formulation for sparse PCA using semidefinite programming.
SIAM review, 49(3):434, 2007.

[14] F. De La Torre and M. Black.
A framework for robust subspace learning.
International Journal of Computer Vision, 54(1):117–142, 2003.

[15] S. Dhanasekaran, T. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. Pienta, M. Rubin, and A. Chinnaiyan.
Delineation of prognostic biomarkers in prostate cancer.
Nature, 412(6849):822–826, 2001.

[16] S. Draghici, P. Khatri, A. Eklund, and Z. Szallasi.
Reliability and reproducibility issues in DNA microarray measurements.
TRENDS in Genetics, 22(2):101–109, 2006.

[17] J. Dreyfuss, M. Johnson, and P. Park.
Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers.
Molecular Cancer, 8(1):71, 2009.

[18] R. Edgar, M. Domrachev, and A. Lash.
Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.
Nucleic acids research, 30(1):207, 2002.

[19] B. Efron and R. Tibshirani.
On testing the significance of sets of genes.
The Annals of Applied Statistics, 1(1):107–129, 2007.

[20] G. Eichler, S. Huang, and D. Ingber.
Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles.
Bioinformatics, 19(17):2321, 2003.

[21] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany.
Outcome signature genes in breast cancer: is there a unique set?
Bioinformatics, 21(2):171, 2005.

[22] M. Emblom-Callahan, M. Chhina, O. Shlobin, S. Ahmad, E. Reese, E. Iyer, D. Cox, R. Brenner, N. Burton, G. Grant, et al.
Genomic Phenotype of Non-cultured Pulmonary Fibroblasts in Idiopathic Pulmonary Fibrosis.
Genomics, 2010.

[23] H. Eysenck.
Systematic reviews: Meta-analysis and its problems.
Bmj, 309(6957):789, 1994.

[24] M. Fischler and R. Bolles.
Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography.
Communications of the ACM, 24(6):381–395, 1981.

[25] R. Fisher.
Question 14: Combining independent tests of significance.
American Statistician, 2(5):30–30J, 1948.

[26] B. Flury.
Common principal components in k groups.
Journal of the American Statistical Association, 79(388):892–898, 1984.

[27] W. Freije, F. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. Liau, P. Mischel, and S. Nelson.
Gene expression profiling of gliomas strongly predicts survival.
Cancer research, 64(18):6503, 2004.

[28] E. Garrett-Mayer, G. Parmigiani, X. Zhong, L. Cope, and E. Gabrielson.
Cross-study validation and combined analysis of gene expression microarray data.
Biostatistics, 9(2):333–354, Apr 2008.

[29] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al.
Bioconductor: open software development for computational biology and bioinformatics.
Genome biology, 5(10):R80, 2004.

[30] D. Ghosh, T. Barette, D. Rhodes, and A. Chinnaiyan.
Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.
Functional & integrative genomics, 3(4):180–188, 2003.

[31] R. Gnanadesikan and J. Kettenring.
Robust estimates, residuals, and outlier detection with multiresponse data.
Biometrics, 28(1):81–124, 1972.

[32] F. Gómez and R. Rodriguez-Roisin.
Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines for chronic obstructive pulmonary disease.
Current opinion in pulmonary medicine, 8(2):81, 2002.

[33] J. Gower.
Multivariate analysis and multidimensional geometry.
The Statistician, 17(1):13–28, 1967.

[34] L. Gravendeel, M. Kouwenhoven, O. Gevaert, J. de Rooi, A. Stubbs, J. Duijm, A. Daemen, F. Bleeker, L. Bralten, N. Kloosterhof, et al.
Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology.
Cancer research, 69(23):9065, 2009.

[35] R. Grutzmann, H. Boriss, O. Ammerpohl, J. Luttges, H. Kalthoff, H. Schackert, G. Kloppel, H. Saeger, and C. Pilarsky.
Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes.
Oncogene, 24:5079–5088, 2005.

[36] Y. Guo, T. Hastie, and R. Tibshirani.
Regularized linear discriminant analysis and its application in microarrays.
Biostatistics, 8(1):86, 2007.

[37] F. Hong and R. Breitling.
A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.
Bioinformatics, 24(3):374, 2008.

[38] F. Hong, R. Breitling, C. McEntee, B. Wittner, J. Nemhauser, and J. Chory.
RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.
Bioinformatics, 22(22):2825, 2006.

[39] H. Hotelling.
The relations of the newer multivariate statistical methods to factor analysis, Brit.
J. Statist. Psychol, 10:69–79, 1957.

[40] P. Huber.
Projection pursuit.
The annals of Statistics, 13(2):435–475, 1985.

[41] P. Huber, E. Ronchetti, and E. Corporation.
Robust statistics, volume 1.
Wiley Online Library, 1981.

[42] M. Hubert and S. Engelen.
Robust PCA and classification in biosciences.
Bioinformatics, 20(11):1728, 2004.

[43] M. Hubert and P. Rousseeuw.
ROBPCA: a new approach to robust principal component analysis.
Technometrics, pages 64–79, 2005.

[44] J. Ioannidis, D. Allison, C. Ball, I. Coulibaly, X. Cui, A. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, et al.
Repeatability of published microarray gene expression analyses.
Nature genetics, 41(2):149–155, 2008.

[45] R. Irizarry, B. Bolstad, F. Collin, L. Cope, B. Hobbs, and T. Speed.
Summaries of Affymetrix GeneChip probe level data.
Nucleic acids research, 31(4):e15, 2003.

[46] I. Jolliffe.
Principal component analysis.
2002.

[47] I. Jolliffe, N. Trendafilov, and M. Uddin.
A modified principal component technique based on the LASSO.
Journal of Computational and Graphical Statistics, 12(3):531–547, 2003.

[48] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre.
Generalized power method for sparse principal component analysis.
The Journal of Machine Learning Research, 11:517–553, 2010.

[49] Y. Jung, M. Oh, D. Shin, S. Kang, and H. Oh.
Identifying Differentially Expressed Genes in Meta-Analysis via Bayesian Model-Based Clustering.
Biometrical Journal, 48(3):435–450, 2006.

[50] D. D. Kang and G. C. Tseng.
Metaqc: Quantitative quality assessment for inclusion/exclusion criteria of genomic meta-analysis.
Manuscript in preparation.

[51] Q. Ke and T. Kanade.
Robust L Norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming.
2005.

[52] M. Kendall and A. Stuart.
The Advanced Theory of Statistics, London: Charles Griffin & Co.
Ltd., paragraph, 33, 1973.

[53] M. G. Kendall.
A course in multivariate analysis.
Hafner Pub. Co. (New York), 1957.

[54] D. Kim, H. Collard, and T. King Jr.
Classification and natural history of the idiopathic interstitial pneumonias.
In *Proceedings of the American Thoraic Society*, volume 3, page 285. Am Thoracic Soc, 2006.

[55] K. Konishi, K. Gibson, K. Lindell, T. Richards, Y. Zhang, R. Dhir, M. Bisceglia, S. Gilbert, S. Yousem, J. Song, et al.
Gene expression profiles of acute exacerbations of
idiopathic pulmonary fibrosis.
American journal of respiratory and critical care medicine, 180(2):167, 2009.

[56] W. Krzanowski.
Between-groups comparison of principal components.
Journal of the American Statistical Association, 74(367):703–707, 1979.

[57] W. Krzanowski.
Principal component analysis in the presence of group structure.
Applied Statistics, 33(2):164–168, 1984.

[58] J. Lapointe, C. Li, J. Higgins, M. Van De Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, et al.
Gene expression profiling identifies clinically relevant subtypes of prostate cancer.
Proceedings of the National Academy of Sciences of the United States of America, 101(3):811, 2004.

[59] O. Larsson, D. Diebold, D. Fan, M. Peterson, R. Nho, P. Bitterman, and C. Henke.
Fibrotic myofibroblasts manifest genome-wide derangements of translational
control.
PLoS One, 3(9):3220, 2008.

[60] M. Lee, H. Shen, J. Huang, and J. Marron.
Biclustering via Sparse Singular Value Decomposition.
Biometrics, 2010.

[61] G. Li and Z. Chen.
Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo.
Journal of the American Statistical Association, 80(391):759–766, 1985.

[62] K. Li, M. Yan, and S. Yuan.
A simple statistical model for depicting the cdc 15-synchronized yeast cell-cycle regulated gene expression data.
Statistica Sinica, 12(1):141–158, 2002.

[63] L. Liu, D. Hawkins, S. Ghosh, and S. Young.
Robust singular value decomposition analysis of microarray data.
Proceedings of the National Academy of Sciences of the United States of America, 100(23):13167, 2003.

[64] S. Lu, J. Li, C. Song, K. Shen, and G. Tseng.
Biomarker detection in the integration of multiple multi-class genomic studies.
Bioinformatics, 26(3):333, 2010.

[65] H. Mann and D. Whitney.
On a test of whether one of two random variables is stochastically larger than the other.
The Annals of Mathematical Statistics, 18(1):50–60, 1947.

[66] D. Mannino, D. Homa, L. Akinbami, E. Ford, and S. Redd.
Chronic obstructive pulmonary disease surveillance-United States, 1971-2000.
Respiratory care, 47(10):1184–1199, 2002.

[67] B. Moghaddam, Y. Weiss, and S. Avidan.
Spectral bounds for sparse PCA: Exact and greedy algorithms.
Advances in Neural Information Processing Systems, 18:915, 2006.

[68] M. Mulligan, I. Ponomarev, R. Hitzemann, J. Belknap, B. Tabakoff, R. Harris, J. Crabbe, Y. Blednov, N. Grahame, T. Phillips, et al.
Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis.
Proceedings of the National Academy of Sciences of the United States of America, 103(16):6368, 2006.

[69] S. Nanni, C. Priolo, A. Grasselli, M. D'Eletto, R. Merola, F. Moretti, M. Gallucci, P. De Carli, S. Sentinelli, A. Cianciulli, et al.
Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer.
Molecular Cancer Research, 4(2):79, 2006.

[70] B. Neuenschwander and B. Flury.
Common principal components for dependent random vectors.
Journal of multivariate analysis, 75(2):163–183, 2000.

[71] B. North, D. Curtis, and P. Sham.
A note on the calculation of empirical P values from Monte Carlo procedures.
American journal of human genetics, 72(2):498, 2003.

[72] S. Oh, D. Kang, G. Brock, and G. Tseng.
Biological impact of missing-value imputation on downstream analyses of gene expression profiles.
Bioinformatics, 27(1):78, 2011.

[73] A. Olson, J. Swigris, D. Lezotte, J. Norris, C. Wilson, and K. Brown.
Mortality from pulmonary fibrosis increased in the United States from 1992 to 2003.
American journal of respiratory and critical care medicine, 176(3):277, 2007.

[74] A. Owen.
Pearson's Test in a Large Scale Multiple Meta-Analysis.
2007.

[75] A. Pardo, K. Gibson, J. Cisneros, T. Richards, Y. Yang, C. Becerril, S. Yousem, I. Herrera, V. Ruiz, M. Selman, et al.
Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis.
PLoS medicine, 2(9):891, 2005.

[76] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. Garcia Lara, E. Holloway, M. Kapushesky, et al.
ArrayExpress–a public repository for microarray gene expression data at the EBI.
Nucleic acids research, 33(Database Issue):D553, 2005.

[77] G. Parmigiani, E. S. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson.
A cross-study comparison of gene expression studies for the molecular classification of lung cancer.
Clin Cancer Res, 10(9):2922–2927, May 2004.

[78] B. Paugh, C. Qu, C. Jones, Z. Liu, M. Adamowicz-Brice, J. Zhang, D. Bax, B. Coyle, J. Barrow, D. Hargrave, et al.
Integrated molecular genetic profiling of pediatric high-grade gliomas reveals key differences with the adult disease.
Journal of Clinical Oncology, 28(18):3061, 2010.

[79] L. Petalidis, A. Oulas, M. Backlund, M. Wayland, L. Liu, K. Plant, L. Happerfield, T. Freeman, P. Poirazi, and V. Collins.
Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data.
Molecular cancer therapeutics, 7(5):1013, 2008.

[80] H. Phillips, S. Kharbanda, R. Chen, W. Forrest, R. Soriano, T. Wu, A. Misra, J. Nigro, H. Colman, L. Soroceanu, et al.
Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.
Cancer Cell, 9(3):157–173, 2006.

[81] S. Pierrou, P. Broberg, R. O'donnell, K. Pawlowski, R. Virtala, E. Lindqvist, A. Richter, S. Wilson, G. Angco, S. Moller, et al.
Expression of genes involved in oxidative stress responses in airway epithelial cells of COPD smokers.
American journal of respiratory and critical care medicine, pages 200607–931OCv1, 2006.

[82] W. Qiu and H. Joe.
Generation of random clusters with specified degree of separation.
Journal of classification, 23(2):315–334, 2006.

[83] R Development Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2010.
ISBN 3-900051-07-0.

[84] A. Ramasamy, A. Mondry, C. Holmes, and D. Altman.
Key issues in conducting a meta-analysis of gene expression microarray datasets.
PLoS Med, 5(9):e184, 2008.

[85] S. Raychaudhuri, J. Stuart, and R. Altman.
Principal components analysis to summarize microarray experiments: application to sporulation time series.
In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 455. NIH Public Access, 2000.

[86] D. Rhodes, T. Barrette, M. Rubin, D. Ghosh, and A. Chinnaiyan.
Meta-Analysis of Microarrays.
Cancer Research, 62(15):4427, 2002.

[87] D. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. Chinnaiyan.
Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.
Proceedings of the national academy of sciences of the United States of America, 101(25):9309, 2004.

[88] J. Sammon Jr.
A Nonlinear Mapping for Data Structure Analysis.
IEEE Transactions on Computers, 18(5), 1969.

[89] A. Sboner, F. Demichelis, S. Calza, Y. Pawitan, S. Setlur, Y. Hoshida, S. Perner, H. Adami, K. Fall, L. Mucci, et al.
Molecular sampling of prostate cancer: a dilemma for predicting disease progression.
BMC Medical Genomics, 3(1):8, 2010.

[90] M. Selman, G. Carrillo, A. Estrada, M. Mejia, C. Becerril, J. Cisneros, M. Gaxiola, R. Pérez-Padilla, C. Navarro, T. Richards, et al.
Accelerated variant of idiopathic pulmonary fibrosis: clinical behavior and gene expression pattern.
PLoS One, 2(5):e482, 2007.

[91] M. Selman, A. Pardo, L. Barrera, A. Estrada, S. Watson, K. Wilson, N. Aziz, N. Kaminski, and A. Zlotnik.
Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis.
American journal of respiratory and critical care medicine, pages 200504–644OCv1, 2005.

[92] A. Sharov, D. Dudekula, and M. Ko.
A web-based tool for principal component and significance analysis of microarray data.
Bioinformatics, 21(10):2548, 2005.

[93] H. Shen and J. Huang.
Sparse principal component analysis via regularized low rank matrix approximation.
Journal of multivariate analysis, 99(6):1015–1034, 2008.

[94] K. Shen and G. Tseng.
Meta-analysis for pathway enrichment analysis when combining multiple microarray studies.
Bioinformatics, 2010.

[95] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. Matese, S. Dwight, M. Kaloper, S. Weng, H. Jin, C. Ball, et al.
The Stanford microarray database.
Nucleic Acids Research, 29(1):152, 2001.

[96] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, et al.
Gene expression correlates of clinical prostate cancer behavior.
Cancer cell, 1(2):203–209, 2002.

[97] D. Smith, P. Sætrom, O. Snøve, C. Lundberg, G. Rivas, C. Glackin, and G. Larson.
Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation.
BMC bioinformatics, 9(1):63, 2008.

[98] A. T. Society.
Idiopathic pulmonary fibrosis: Diagnosis and treatment . international consensus statement.
Am. J. Respir. Crit. Care Med., 161(2):646–664, 2000.

[99] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher.
Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.
Molecular biology of the cell, 9(12):3273, 1998.

[100] A. Spira, J. Beane, V. Pinto-Plata, A. Kadar, G. Liu, V. Shah, B. Celli, and J. Brody.
Gene expression profiling of human lung tissue from smokers with severe emphysema.
American journal of respiratory cell and molecular biology, 31(6):601–610, 2004.

[101] J. Stevens and R. Doerge.
Combining Affymetrix microarray results.
BMC bioinformatics, 6(1):57, 2005.

[102] S. Stouffer, E. Suchman, L. DeVinney, S. Star, and R. Williams Jr.
The American soldier: adjustment during army life.(Studies in social psychology in World War II, Vol. 1.).
1949.

[103] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al.
Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.
Proceedings of the National Academy of Sciences of the United States of America, 102(43):15545, 2005.

[104] L. Sun, A. Hui, Q. Su, A. Vortmeyer, Y. Kotliarov, S. Pastorino, A. Passaniti, J. Menon, J. Walling, R. Bailey, et al.
Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain.
Cancer cell, 9(4):287–300, 2006.

[105] P. Tan, T. Downey, E. Spitznagel Jr, P. Xu, D. Fu, D. Dimitrov, R. Lempicki, B. Raaka, and M. Cam.
Evaluation of gene expression measurements from commercial microarray platforms.
Nucleic acids research, 31(19):5676, 2003.

[106] S. Theodoridis and K. Koutroumbas.
Pattern Recognition, Third Edition.
Academic Press, Inc., Orlando, FL, USA, 2006.

[107] S. Tomlins, R. Mehra, D. Rhodes, X. Cao, L. Wang, S. Dhanasekaran, S. Kalyana-Sundaram, J. Wei, M. Rubin, K. Pienta, et al.
Integrative molecular concept modeling of prostate cancer progression.
Nature genetics, 39(1):41–51, 2006.

[108] G. Tseng, M. Oh, L. Rohlin, J. Liao, and W. Wong.
Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.
Nucleic Acids Research, 29(12):2549, 2001.

[109] S. Varambally, J. Yu, B. Laxman, D. Rhodes, R. Mehra, S. Tomlins, R. Shah, U. Chandran, F. Monzon, M. Becich, et al.
Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression.
Cancer cell, 8(5):393–406, 2005.

[110] L. Vuga, A. Ben-Yehudah, E. Kovkarova-Naumovski, T. Oriss, K. Gibson, C. Feghali-Bostwick, and N. Kaminski.
WNT5A is a regulator of fibroblast proliferation and resistance to apoptosis.
American journal of respiratory cell and molecular biology, 41(5):583–589, 2009.

[111] T. Wallace, R. Prueitt, M. Yi, T. Howe, J. Gillespie, H. Yfantis, R. Stephens, N. Caporaso, C. Loffredo, and S. Ambs.
Tumor immunobiological differences in prostate cancer between African-American and European-American men.
Cancer research, 68(3):927, 2008.

[112] J. Welsh, L. Sapinoso, A. Su, S. Kern, J. Wang-Rodriguez, C. Moskaluk, H. Frierson, and G. Hampton.
Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer.
Cancer Research, 61(16):5974, 2001.

[113] P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, et al.
Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures.
Breast Cancer Res, 10(4):R65, 2008.

[114] D. Witten and R. Tibshirani.
A framework for feature selection in clustering.
Journal of the American Statistical Association, 105(490):713–726, 2010.

[115] D. Witten, R. Tibshirani, and T. Hastie.
A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.
Biostatistics, 10(3):515, 2009.

[116] P. Woodruff, H. Boushey, G. Dolganov, C. Barker, Y. Yang, S. Donnelly, A. Ellwanger, S. Sidhu, T. Dao-Pick, C. Pantoja, et al.
Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids.
Proceedings of the National Academy of Sciences, 104(40):15858, 2007.

[117] P. Woodruff, L. Koth, Y. Yang, M. Rodriguez, S. Favoreto, G. Dolganov, A. Paquet, and D. Erle.
A distinctive alveolar macrophage activation state induced by cigarette smoking.
American journal of respiratory and critical care medicine, pages 200505–686OCv1, 2005.

[118] R. Yamanaka, T. Arao, N. Yajima, N. Tsuchiya, J. Homma, R. Tanaka, M. Sano, A. Oide, M. Sekijima, and K. Nishio.
Identification of expressed genes characterizing long-term survival in malignant glioma patients.
Oncogene, 25(44):5994–6002, 2006.

[119] I. Yang, L. Burch, M. Steele, J. Savov, J. Hollingsworth, E. McElvania-Tekippe, K. Berman, M. Speer, T. Sporn, K. Brown, et al.
Gene expression profiling of familial and sporadic cases of interstitial pneumonia.
American Journal of Respiratory and Critical Care Medicine, page 200601, 2006.

[120] K. Yeung and W. Ruzzo.
Principal component analysis for clustering gene expression data.
Bioinformatics, 17(9):763, 2001.

[121] W. Youden.
Index for rating diagnostic tests.
Cancer, 3(1):32–35, 1950.

[122] Y. Yu, D. Landsittel, L. Jing, J. Nelson, B. Ren, L. Liu, C. McDonald, R. Thomas, R. Dhir, S. Finkelstein, et al.
Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy.
Journal of Clinical Oncology, 22(14):2790, 2004.

[123] H. Zou, T. Hastie, and R. Tibshirani.
Sparse principal component analysis.
Journal of computational and graphical statistics, 15(2):265–286, 2006.

[124] F. Zuo, N. Kaminski, E. Eugui, J. Allard, Z. Yakhini, A. Ben-Dor, L. Lollini, D. Morris, Y. Kim, B. DeLustro, et al.
Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans.
Proceedings of the National Academy of Sciences, 99(9):6292, 2002.